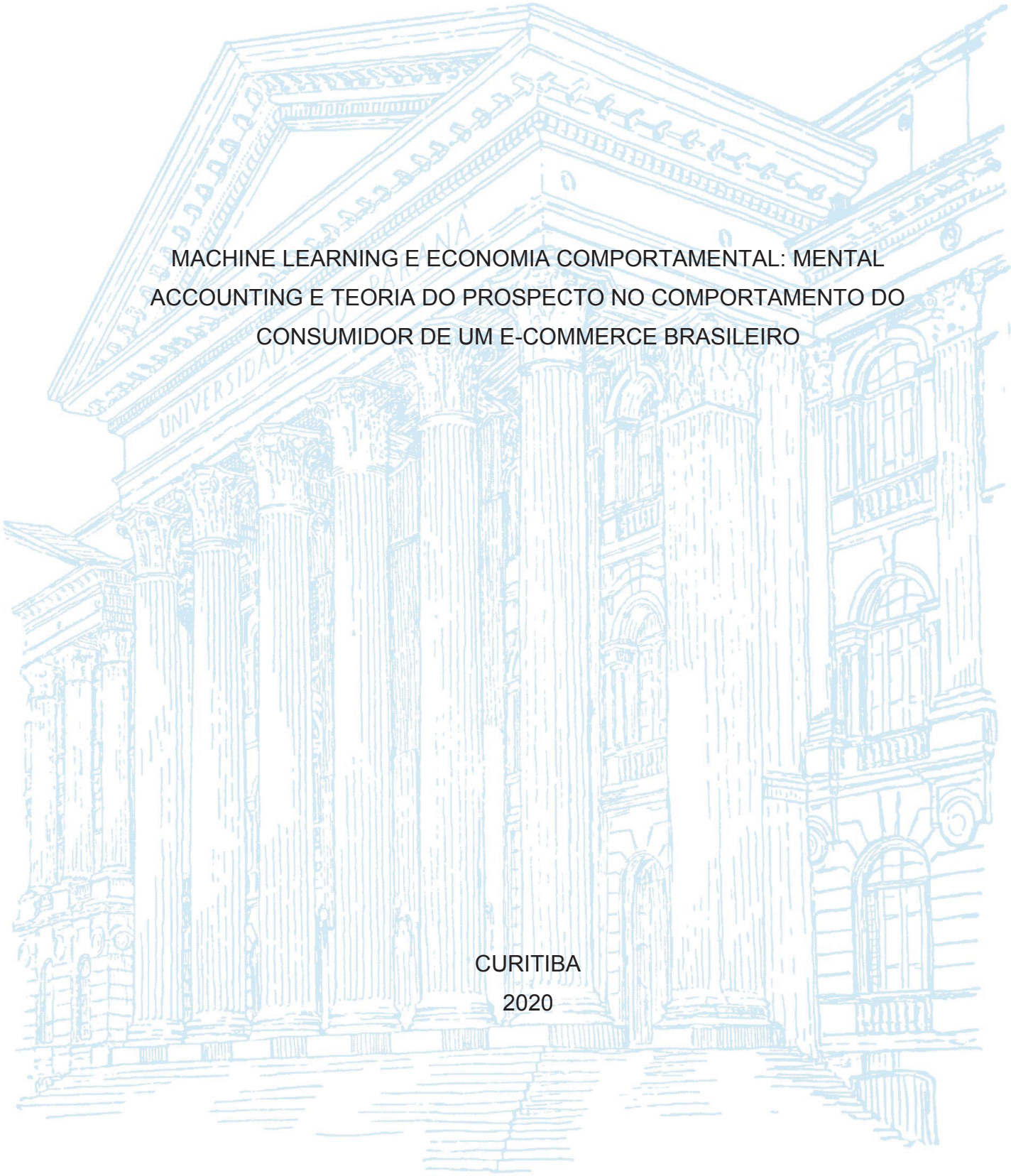
UNIVERSIDADE FEDERAL DO PARANÁ

DIEGO LIBERATO SOUZA



MACHINE LEARNING E ECONOMIA COMPORTAMENTAL: MENTAL ACCOUNTING E TEORIA DO PROSPECTO NO COMPORTAMENTO DO CONSUMIDOR DE UM E-COMMERCE BRASILEIRO

CURITIBA

2020

DIEGO LIBERATO SOUZA

MACHINE LEARNING E ECONOMIA COMPORTAMENTAL:MENTAL ACCOUNTING E TEORIA DO PROSPECTO NO COMPORTAMENTO DO CONSUMIDOR DE UM E-COMMERCE BRASILEIRO

Dissertação apresentada ao Curso de Pós-Graduação em Desenvolvimento Econômico, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Desenvolvimento Econômico.

Orientadora: Prof. Dr(a). Adriana Sbicca Fernandes

CURITIBA

2020

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS – SIBI/UFPR COM DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário: Eduardo Silveira – CRB 9/1921

Souza, Diego Liberato
Machine learning e economia comportamental: mental accounting e teoria do prospecto no comportamento do consumidor de um e-commerce brasileiro / Diego Liberato Souza.- 2020.
138 p.

Dissertação (Mestrado) - Universidade Federal do Paraná. Programa de Pós-Graduação em Desenvolvimento Econômico, do Setor de Ciências Sociais Aplicadas.

Orientadora: Adriana Sbicca Fernandes.

Defesa: Curitiba, 2020.

1. Economia. 2. Comportamento do consumidor. I. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós-Graduação em Desenvolvimento Econômico. II. Fernandes, Adriana Sbicca. III. Título.

CDD 330

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIENCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO DESENVOLVIMENTO ECONÔMICO - 40001016024P0

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em DESENVOLVIMENTO ECONÔMICO da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **DIEGO LIBERATO SOUZA** intitulada: **MACHINE LEARNING E ECONOMIA COMPORTAMENTAL: MENTAL ACCOUNTING E TEORIA DO PROSPECTO NO COMPORTAMENTO DO CONSUMIDOR DE UM E-COMMERCE BRASILEIRO**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 21 de Fevereiro de 2020.

Assinatura Eletrônica
04/05/2020 17:49:37.0
ADRIANA SBICCA FERNANDES
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
04/05/2020 15:45:04.0
DENISE FUKUMI TSUNODA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
07/05/2020 14:48:33.0
HERMES YUKIO HIGACHI
Avaliador Externo (UNIVERSIDADE ESTADUAL DE PONTA GROSSA)

Dedico este trabalho aos meus pais, Sergio e Raquel, ao meu irmão e irmã, Pedro e Carol, à Bruna e a todos aqueles que me apoiaram nesta jornada.

AGRADECIMENTOS

Agradeço aos meus pais por todo apoio que me deram em relação aos estudos desde o início e todo incentivo que me ofereceram ao longo dessa jornada. Agradeço ao meu irmão, com quem sempre pude discutir este trabalho e que sempre esteve disponível para me ouvir. Agradeço à Bruna, por estar comigo nos momentos mais difíceis, por toda compreensão, ajuda e carinho que me ofereceu. Agradeço a minha orientadora, Adriana Sbicca, por ter me dado a oportunidade de trazer este tema que tanto gosto em minha dissertação, por ajudar em tudo que foi necessário e sempre ser compreensiva.

“The scientist does not study nature because it is useful to do so. He studies it because he takes pleasure in it, and he takes pleasure in it because it is beautiful. If nature were not beautiful it would not be worth knowing, and life would not be worth living”

(Henri Poincaré, *Science and Method*, 2007)

RESUMO

Este trabalho busca encontrar vieses cognitivos em consumidores de e-commerce utilizando Machine Learning e uma abordagem preditiva em uma base de dados real de uma empresa de E-commerce brasileira. A maior parte dos trabalhos que buscam avaliar comportamentos de consumidor se baseiam em técnicas de correlação (Larasati et al., 2012). Este trabalho se diferencia por utilizar os algoritmos de Decision Tree, Random Forests e ANN, além de Regressão Linear Múltipla e Regressão Logística. Dois ensaios são apresentados, sendo que no primeiro ensaio é avaliado o efeito do Mental Accouting denominado de Relative Thinking (Azar, 2007) em que consumidores tendem a pagar fretes maiores para produtos mais caros. O segundo ensaio analisa as variáveis relevantes para a satisfação do consumidor de e-commerce (Qualidade de Informação, Qualidade de Entrega, Preço do Produto e Qualidade de Serviço), testa a Teoria da Expectativa e Desconfirmação (Oliver, 1993) e a assimetria do impacto da Teoria da Expectativa e Desconfirmação na satisfação do consumidor (Youjae Yi e Suna La, 2003), indicando concordância com a Teoria do Prospecto e da Função Valor de Kahneman e Tversky (1979;1992).

Palavras-chave: Machine Learning. Mental Accounting. Teoria do Prospecto. Desconfirmação. Comportamento do Consumidor.

ABSTRACT

This work seeks to find cognitive biases in e-commerce consumers using Machine Learning and a predictive approach in a real database of a Brazilian E-commerce company. Most studies that seek to assess consumer behavior are based on correlation techniques (Larasati et al., 2012). This work is different because it uses Decision Tree, Random Forests and ANN algorithms, in addition to Multiple Linear Regression and Logistic Regression. Two essays are presented, the first essay evaluating the effect of Mental Accouting called Relative Thinking (Azar, 2007) where consumers tend to pay higher freight for more expensive products. The second essay analyzes the relevant variables for e-commerce consumer satisfaction (Information Quality, Delivery Quality, Product Price and Service Quality), testing Expectation and Disconfirmation Theory (Oliver, 1993) and asymmetry of impact of the Theory of Expectation and Disconfirmation on consumer satisfaction (Youjae Yi and Suna La, 2003), indicating the existence of the Prospect Theory and the Value Function by Kahneman and Tversky (1979; 1992).

Keywords: Machine Learning. Mental Accounting. Prospect Theory. Disconfirmation. Customer behavior.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O presente trabalho busca compreender o comportamento do consumidor e identificar vieses cognitivos através da utilização de modelos de Machine Learning. O trabalho se divide em dois ensaios, sendo que no primeiro ensaio é analisado um tipo de específico de Mental Accounting identificado por Thaler (1980), Kahneman e Tversky (1981), Ranyard e Abdel-Nabi (1993), Frisch (1993) e Moon et al. (1999) no qual os consumidores percebem valores financeiros de maneira relativa e não de maneira absoluta, conforme prevê a Teoria Econômica Tradicional.

Esse fenômeno foi definidor por Azar (2007) como Relative Thinking e é descrito como uma tendência cognitiva na qual as pessoas tendem a considerar as diferenças de preços relativas em detrimento dos preços absolutos. Assim, no primeiro ensaio é averiguado se este tipo de Mental Accounting pode ser encontrado no comportamento do consumidor em relação ao frete, fazendo com que o consumidor não olhe o valor financeiro absoluto do frete, mas sim o valor do frete em relação ao preço do produto, de modo que ele esteja mais disposto a pagar mais pelo frete quando o preço do produto é maior. Além disso foi analisado se o impacto do Relative Thinking varia conforme a categoria do produto comprado.

O segundo ensaio se debruça sobre a satisfação do consumidor, com intuito de entender quais são os principais componentes que a afetam. Diversas pesquisas levantaram que variáveis como Qualidade do Produto, Qualidade de informação e Qualidade da Entrega afetam fortemente a satisfação do consumidor (Brown e Jayakody ,2008; Lin, 2007; DeLone e McLean, 2003), assim como o preço do produto (Cao e Gruca, 2004). No presente trabalho os impactos de todos esses componentes na satisfação do consumidor são analisados, assim como, se o modelo de Expectativa e Desconfirmação (Oliver, 1993) afeta a satisfação do consumidor. De acordo com esse modelo, a satisfação do consumidor não depende somente da performance do produto per si, mas também da expectativa prévia que o consumidor tinha em relação ao produto.

Por fim, é analisado o impacto assimétrico da Desconfirmação na satisfação do consumidor encontrado por Youjae Yi e Suna La (2003), que sustenta que a Desconfirmação Negativa (performance abaixo da expectativa) tende a ter maior impacto na saitsfação do consumidor do que um mesmo montante de

Desconfirmação Positiva (performance acima da expectativa), apontando para o fenômeno de Aversão a Perda presente na Teoria do Prospecto (Kahneman e Tversky, 1979; 1992). Evidenciando o fenômeno da Teoria do Prospecto e o impacto assimétrico da Desconfirmação, é analisado se a relação entre Desconfirmação e Satisfação do consumidor satisfaz às condições de Sensibilidade Decrescente (para ganhos e perdas) e de Aversão a Perdas, conforme teorizado por Kahneman e Tversky (1979, 1992). Por fim, é investigado se a Sensibilidade Decrescente e a Aversão a Perda varia conforme o método de pagamento, Estado do consumidor e categoria do produto.

A maior parte dos trabalhos que investigam os mesmos comportamentos citados anteriormente utilizam dados de survey e técnicas de correlação (Ho et al., 2017). Além de utilizar dados reais de consumidores de e-commerce, este trabalho se mostra relevante por expandir o arsenal de técnicas que podem ser analisadas para compreender o comportamento do consumidor, especificamente na identificação de vieses cognitivos.

Baseado em James et al. (2013), existe um trade-off entre intepretabilidade e acurácia, isto é, algoritmos mais interpretáveis tendem a não ser acurados e algoritmos acurados tendem a ser pouco interpretáveis. Desta maneira, o presente trabalho escolheu a utilização tanto de algoritmos interpretáveis (Regressão Linear Múltipla, Regressão Logística, Árvore de Decisão e Non-Linear Least Square), quanto de algoritmos que tendem a ser mais acurados e menos interpretáveis (Random Forest e Artificial Neural Networks) para analisar o comportamento do consumidor. Além disso, foram utilizadas técnicas para aumentar a interpretabilidade dos algoritmos Random Forest e Artificial Neural Networks, como por exemplo Permutation Importance (Fisher, Rudin e Dominici, 2018), H-Statistic (Friedman e Popescu, 2008), Individual Conditional Expectation (Goldstein et al., 2017) e Shapley Values (Štrumbelj e Kononenko, 2014).

2. *ENSAIO 1: MENTAL ACCOUNTING* E *RELATIVE THINKING*: EVIDÊNCIAS NO COMPORTAMENTO DO CONSUMIDOR DE *E-COMMERCE* UTILIZANDO *MACHINE LEARNING*

RESUMO:

Este trabalho investiga a presença de um *Mental Accounting* baseado em Thaler (1980) e Tversky e Kahneman (1981), no qual as pessoas estariam mais dispostas a pagar um frete maior quando o preço do produto é maior por avaliarem o valor do frete em relação a um ponto de referência (preço do produto). A partir da base de dados de uma empresa de *E-commerce* brasileira, foi utilizado *Machine Learning* para analisar se este fenômeno cognitivo realmente ocorre e se varia de acordo com o estado do consumidor e a categoria do produto.

Palavras-chave: *Mental Accounting*, *Machine Learning*, *E-Commerce*

ABSTRACT:

This work investigates the presence of a Mental Accounting based on Thaler (1980) and Tversky and Kahneman (1981), in which people would be more willing to pay a higher freight when the price of the product is higher because they evaluate the value of freight in relation to a reference point (product price). From the database of a Brazilian E-commerce company, Machine Learning was used to analyze whether this cognitive phenomenon really occurs and if it varies according to the product category.

Palavras-chave: *Mental Accounting*, *Machine Learning*, *E-Commerce*

2.1. INTRODUÇÃO

De acordo com Richard Thaler, em seu artigo "*Mental Accounting Matters"* (1999*), Mental Accounting* se refere às operações mentais que as pessoas tendem a utilizar para organizar e avaliar diversos tipos de atividades financeiras, sendo que este fenômeno já foi abordado pelo próprio autor em diversos outros trabalhos (Thaler, 1985; Thaler, 1990; Thaler, 1992). No presente trabalho é discutido um tipo específico de *Mental Accounting*, primeiramente encontrado por Thaler (1980). O autor observou que as pessoas estavam mais dispostas a salvar $5 em um rádio que custa $25 do que uma televisão de $500, indicando que o valor salvo de $5 não é o único valor levado em consideração na operação mental realizada pelos participantes, mas também é considerado o preço de referência do produto, sendo que $5 em relação a $25 aparenta ser maior do que $5 em relação a $500.

Este mesmo efeito foi observado em um clássico experimento de Kahneman e Tversky (1981), em que os autores pediram para os participantes escolherem se estariam dispostos a dirigir 20 minutos para comprar uma calculadora e uma jaqueta, economizando $5 na calculadora. Mais especificamente, duas situações distintas foram oferecidas para dois grupos: na primeira a calculadora custava $15 e a jaqueta $125 e na segunda a calculadora custava $125 e a jaqueta $15. Este mesmo experimento foi replicado por diversos autores (Ranyard e Abdel-Nabi, 1993; Frisch, 1993; Moon et al., 1999) e evidenciam o mesmo fenômeno analisado por Thaler (1980): a maior parte das pessoas prefere economizar $5 na calculadora quando a calculadora custa $15 e a jaqueta custa $125.

Azar (2006) denomina este tipo de Mental Accounting como Relative Thinking e o define como um fenômeno cognitivo no qual as pessoas tendem a considerar as diferenças de preços relativas em detrimento dos preços absolutos, sendo que pela Teoria da Escolha Racional apenas as diferenças absolutas de preços deveriam importar na decisão. Azar (2007) desenvolve um modelo para explicar o funcionamento do *Relative Thinking* e propõe que a intensidade deste fenômeno varia entre as pessoas.

Partindo desta discussão, o presente trabalho investiga se este fenômeno se encontra presente nas decisões tomadas pelos consumidores no valor do frete pago em relação ao preço do produto ao comprar pela internet. De acordo com o *Relative Thinking*, o que se espera é que os consumidores estejam dispostos a economizar

mais em frete quando o preço do produto é menor do que quando o preço do produto é maior, isto é, os consumidores considerariam o preço relativo do frete em relação ao preço do produto e não o preço absoluto do frete. Além disso, é analisado se a intensidade deste *Mental Accounting* varia conforme o estado brasileiro que o consumidor se encontra e a categoria de produto.

O presente trabalho se mostra relevante primeiramente por, diferentemente da maioria dos trabalhos, utilizar dados reais de uma empresa em detrimento de questionários, visto que o único trabalho a fazer isso para analisar *Relative Thinking* foi Hirshman et al. (2018). O segundo motivo pelo qual este trabalho se mostra relevante é analisar este fenômeno sob um contexto novo, isto é, investigar se este viés está presente na decisão do consumidor de pagar um valor de frete maior ou menor e também sob o contexto de *E-commerce*, sendo este fenômeno só foi discutido no trabalho de Da Liang e Chen (2012). A terceira razão da importância deste trabalho é que consegue mostrar que existem diferenças deste viés cognitivo entre os estados brasileiros e entre diferentes categorias de produto. Por fim, a quarta justificativa de relevância deste trabalho se encontra em sua metodologia, visto que a maior parte dos trabalhos utilizam testes de hipótese a partir de questionários, enquanto este trabalho se diferencia por utilizar *Machine Learning*, sendo que tais técnicas se mostram mais eficientes que metodologias tradicionais para analisar comportamentos do consumidor (Garver, 2002).

A seção 2 traz a Revisão Bibliográfica acerca do tema de *Relative Thinking* e da relação deste efeito com custos de transporte. A Seção 3 apresenta a metodologia utilizada, a explicação do funcionamento, assim como das configurações dos algoritmos e as métricas utilizadas para avaliar a performance dos algoritmos. A Seção 4 mostra os resultados obtidos a partir da aplicação destes algoritmos. A Seção 5 traz as conclusões.

2.2. REVISÃO BIBLIOGRÁFICA

2.2.1. MENTAL ACCOUNTING E *RELATIVE THINKING*

O efeito do *Relative Thinking* tem sido fortemente investigado por Azar (2006, 2007, 2008, 2009, 2010, 2011a, 2011b, 2011c, 2013, 2014). A origem deste fenômeno cognitivo remonta aos resultados relacionados ao *Mental Accounting*

encontrado por Thaler (1980), no qual o autor encontrou evidência de que as pessoas estavam mais dispostas a salvar 5 dólares em um rádio que custa 25 dólares, do que 5 dólares em uma televisão de 500 dólares. Essa observação contraria a teoria da maximização da utilidade tradicional, em que se deveria apenas considerar a diferença absoluta de $5 e não esse valor em relação ao preço do Rádio ou da Televisão.

Tversky e Kahneman (1981) realizaram um experimento baseado em Thaler (1980) e Savage (1972) e encontram evidências semelhantes. A partir de um questionário, foi perguntado a dois grupos de participantes se eles estariam dispostos a dirigir 20 minutos para comprar uma jaqueta e uma calculadora, sendo que economizariam $5 na calculadora. A pergunta apresentada por Tversky e Kahneman (1981) pode ser vista abaixo, sendo que os valores em parênteses são os preços dos produtos apresentados para um grupo e os valores em colchetes são os preços dos produtos apresentados para o outro grupo:

"Imagine que você vai comprar uma jaqueta por ($125)[$15], e uma calculadora por ($15)[$125]. O vendedor da calculadora informa que essa mesma calculadora que você quer comprar está a venda por ($10)[$120] em outra filial da loja, localizada a 20 minutos de distância. Você faria esta viagem até a outra loja" (Tversky e Kahneman, 1981, tradução nossa)

Os autores encontraram que quando a calculadora era $15 e a jaqueta $125, 68% das pessoas estavam dispostas a dirigir e quando a calculadora era $125 e a jaqueta $15, 29% das pessoas estavam dispostas. Este resultado está de acordo com o que foi encontrado por Thaler (1980), visto que os consumidores estavam mais dispostos a economizar $5 quando a calculadora custava $15 do que quando custava $125. Tversky e Kahneman (1981) justificam este resultado pelo fato de os valores do desconto serem tomados a partir de um ponto de referência tido como neutro (no caso, o preço do produto do qual será o desconto). Ao comparar com os respectivos pontos de referências, os participantes consideraram $5 pouco quando o preço do produto (ponto de referência) era maior, e estavam dispostos a poupar $5 apenas quando o preço do produto era menor.

Mowen e Mowen (1986) replicaram este experimento e encontraram resultados semelhantes ao que foi encontrado por Tversky e Kahneman (1981),

além disso, fizeram mais dois experimentos para ver se este mesmo fenômeno cognitivo se repetia com decisões relacionadas ao contexto de negócios, aplicando tanto para estudantes quanto para executivos, sendo que o resultado mostrou o mesmo efeito cognitivo observado em Tversky e Kahneman (1981). Ranyard e Abdel-Nabi (1993) encontraram os mesmos resultados variando o preço da jaqueta ao invés da calculadora e Frisch (1993) encontrou o mesmo efeito dos demais experimentos considerando que apenas a jaqueta está sendo comprada. Por fim, Moon et al. (1999) encontraram resultados semelhantes aos demais trabalhos.

Ao longo destes experimentos, o fenômeno cognitivo encontrado nos experimentos foi descrito como uma forma de *Mental Accounting*, mais especificamente uma maneira pela qual as pessoas tendem a enquadrar valores a partir de um valor de referência (Tversky e Kahneman, 1981). Tal efeito foi detectado ao deixar uma informação relevante constante (no caso, o valor do desconto) e alterando o enquadramento da decisão através de uma mudança de contexto irrelevante (Mowen e Mowen, 1986). Azar (2006) denomina este efeito de *Relative* Thinking e o define como um viés no qual as pessoas consideram os preços relativos em detrimento dos preços absolutos, sendo que apenas diferenças absolutas de preços deveriam importar na decisão. A análise das opções oferecidas no experimento realizado por Tversky e Kahneman (1981) não deveria levar em consideração o preço do produto, mas de quanto seria economizado ao dirigir até a outra filial (sendo que em ambos os cenários o valor economizado seria $5).

Segundo Azar (2008), *Mental Accounting* não é o melhor termo para este fenômeno, sendo o termo *Relative Thinking* um termo mais apropriado, primeiramente pelo fato de *Mental Accounting* ser um termo mais amplo e utilizado para descrever diversos tipos de operações mentais específicas que são completamente diferentes entre si e que não se relacionam com o que foi identificado no experimentos citados anteriormente. Além disso, o termo *Relative Thinking* captura melhor a ideia do fenômeno encontrado pela literatura, no qual as pessoas pensam em magnitudes relativas e não absolutas.

Azar (2007) mostra um experimento com um desenho diferente dos demais descritos anteriormente, onde ele pediu aos participantes suporem que vão comprar determinado bem, dizendo a eles qual era o preço do bem na loja. Em seguida ele perguntou aos participantes qual seria o preço máximo para que eles se dispusessem a comprar o mesmo produto em outra loja, sendo que demorariam 20

minutos dirigindo para chegar a essa nova loja. O que ocorreu foi que o valor da diferença de preços aceita pelos participantes para dirigir 20 minutos para comprar o produto mais barato foi de $1,88 (ao comprar uma caneta de $3) para $454,81 (em um produto de $30.000).

Este valor está de acordo com os resultados anteriores e indica, segundo Azar (2007), que pessoas se comportam como se o valor do seu tempo fosse maior quando eles compram um bem mais caro, visto que 20 minutos tem custos diferentes para produtos diferentes, quando teoricamente deveriam ter o mesmo valor independente do produto. Além disso, o resultado indica que o fenômeno de *Relative Thinking* possui intensidade diferente entre as pessoas, já que elas estavam dispostas a aceitar valores diferentes para produtos de mesmo preço. Desta maneira, Azar (2007) propõe um modelo teórico no qual o autor insere os parâmetros α e β, no qual o primeiro captura a extensão do fenômeno cognitivo mostrado nos experimentos anteriores para um determinado consumidor e o segundo captura a extensão deste mesmo fenômeno em um determinado contexto de decisão, levantando desta maneira a hipótese de que diferentes consumidores podem possuir diferentes sensibilidades ao fenômeno encontrado pela literatura.

O modelo desenvolvido por Azar (2007) e a conceituação dada pelos trabalhos anteriores estão intimamente relacionados com o conceito de sensibilidade decrescente, no qual as pessoas tendem a possuir uma sensibilidade decrescente a partir de um ponto de referência, como descrito por Tversky e Kahneman (1981). Contudo, existe outra abordagem para explicar o fenômeno de *Relative Thinking,* apresentada por Bushong et al. (2015), na qual é sustentado que o consumidor não leva em consideração somente sua cesta de consumo (*c*), mas também um conjunto de comparação (*C*). O modelo desenvolvido pelos autores identifica que o conjunto de comparação *C* distorce os pesos relativos que o consumidor atribui aos bens de sua cesta de consumo.

O objetivo do presente trabalho não é discutir qual é a abordagem correta para definir *Relative Thinking*, sendo que aqui será considerado o conceito de *Relative Thinking* relativo à sensibilidade decrescente em relação a um ponto de referência. Azar (2007) defende que esta abordagem do fenômeno cognitivo em questão está associada à lei de Weber, a qual propõe que nossa habilidade em distinguir entre dois estímulos depende de suas diferenças relativas e não de suas diferenças absolutas. Segundo a lei de Weber, se x_1 é o primeiro estímulo e x_2 é o

segundo estímulo, não comparamos os dois a partir da diferença absoluta entre os dois estímulos (x_2 - x_1) mas sim pela sua diferença relativa (x_2 / x_1). De maneira semelhante, o conceito de *Relative Thinking* proposto por Azar (2007) mostra que as pessoas tendem a olhar o preço em comparação com outro preço de referência, isto é, as pessoas consideram as diferenças de preços relativas em detrimento dos preços absolutos.

Apesar das diversas evidências a favor do fenômeno de *Relative Thinking*, o trabalho de Hasting e Shapiro (2013) traz um resultado que vai em direção oposta ao que foi encontrado nos experimentos anteriores. Utilizando microdados de consumo de gasolina através de um modelo de escolha discreta, Hasting e Shapiro (2013) analisaram o comportamento do consumo de gasolina em relação ao consumo da gasolina normal e gasolina *premium* (com maior índice de octanagem) quando o preço da gasolina sobe. O que os autores observaram foi que quando o preço da gasolina sobe as pessoas estão menos dispostas a comprar a gasolina *premium.*

De acordo com o *Relative Thinking*, a diferença relativa de preços da gasolina normal e *premium* pareceria menor quando o preço da gasolina sobe, fazendo com que as pessoas consumissem mais gasolina *premium*, contudo foi observado justamente o contrário. Este resultado sugere que o efeito de *Relative Thinking* não acontece com todos os tipos de produtos, como por exemplo a gasolina.

Hirshman et al. (2018) utilizaram um experimento para observar se o fenômeno de *Relative Thinking* varia com diferentes tipos de produtos. Os autores selecionaram um conjunto de produtos (caneta, clipe de papel, doce, leite, limpador de vidros e gasolina) para ver se os participantes estavam dispostos a comprar a versão *premium* dos mesmos conforme os preços dos produtos normais e *premium* aumentavam. O resultado encontrado foi que, conforme os preços da versão normal e *premium* aumentam, as pessoas estavam mais dispostas a adquirir a versão *premium*, evidenciando o fenômeno de *Relative Thinking*. Contudo este fenômeno não ocorreu para a gasolina, na realidade, conforme os preços aumentaram os participantes se mostraram menos dispostos a adquirir a versão *premium*. Este resultado reforça que o fenômeno de *Relative Thinking* varia conforme o tipo de produto comprado pelo consumidor.

No contexto de *e*-commerce a única evidência deste fenômeno se encontra no trabalho de Da Liang e Chen (2012), em que os autores encontraram evidência

do fenômeno de *Relative Thinking* em leilões online de cupons de hotel. Por fim, a única evidência de *Relative Thinking* retirada de base de dados foi encontrada por Hirshman et al. (2018), em que os autores analisaram a compra de seguros para carros alugados e observaram que as pessoas estão mais dispostas a pagar o seguro quando o preço do carro é mais alto.

2.2.2. *RELATIVE THINKING* E CUSTOS DE TRANSPORTE

Baseado no fenômeno de *Relative Thinking* descrito na seção anterior observa-se que consumidores estão mais dispostos a salvar determinada quantidade de dinheiro em produtos mais baratos do que em produtos mais caros, como mostrado em Thaler (1980) e nos demais experimentos descritos na seção anterior, indicando que os consumidores se esforçam mais para economizar quando a poupança relativa é alta (Azar, 2008), sendo que isso tende a ocorrer com produtos mais baratos. Além disso, baseado em Azar (2007), os experimentos indicam que quanto maior o valor do produto maior é o custo de tempo do consumidor, dado que para produtos mais caros o valor do desconto tem que ser maior para que ele dirija 20 minutos para buscar em outra filial.

Azar (2008) desenvolveu um modelo de diferenciação de localidades com custos de transportes endógenos, isto é, onde o custo de transporte é uma variável endógena. O autor observou que no contexto de diferenciação das localidades, os consumidores se comportam como se o custo de transporte aumentasse de acordo com o preço do produto, sendo que como custo de transporte o autor se refere a ao custo de ir buscar o produto a uma distância maior (e não o valor do frete em si, como geralmente é definido). Logo, baseado nesta modelagem, Azar (2008) observa que a preços maiores os consumidores estariam menos dispostos a ir buscar o produto mais distante e de procurar alternativas mais próximas, visto que o custo de tempo do consumidor (seja o custo de transporte ou custo de procura, isto é, para procurar preços alternativos) é maior conforme o preço do produto. Isso indica que o consumidor não olha o custo de transporte como um valor absoluto, mas de maneira relativa ao preço do produto: quanto maior o preço do produto, menor sua disposição de buscar o produto em um lugar mais distante, semelhante ao que é observado em Tversky e Kahneman (1981), e de procurar alternativas.

Azar (2013) mostra que o fenômeno de *Relative Thinking* explica o quebra cabeça referente a dispersão de preços, no qual foi observado que existe uma correlação entre a dispersão de preços e o preço dos produtos, sendo que quando os preços são maiores a dispersão de preços é maior. Assim, Azar (2013) mostra que este quebra cabeça existe devido ao pressuposto de que o custo de transporte e os custos de procura (isto é, o custo de procurar por alternativas mais próximas ou de menor preço para o mesmo produto) são independentes do preço, contudo, devido ao fenômeno de *Relative Thinking,* as pessoas tendem a se comportar como se o valor de seu tempo fosse maior quando estão comprando um produto mais caro, sendo que desta maneira os custos de procura e de transporte tendem a parecer ser maiores para produtos mais caros e poupanças relativas menores, fazendo com que o consumidor fique indiferente em gastar uma quantidade maior de dinheiro para não ter de procurar ou buscar o produto.

Apesar de não haver nenhuma literatura abordando o fenômeno de *Relative Thinking* em relação ao frete, os resultados sugeridos pelos experimentos de *Relative Thinking* e dos trabalhos de Azar (2008) e Azar (2013) indicam que os custos de procura e de transporte (isto é, de ir buscar o produto) não são levados em consideração como valores absolutos, mas como valores relativos ao preço do produto. Desta maneira, no processo decisório do indivíduo, os custos de procura e transporte são percebidos como maiores quando o preço do produto é maior, fazendo com que eles estejam dispostos a pagar maiores valores de frete para produtos mais caros, evitando custos de procura e transporte. Além disso, baseado na literatura descrita na seção anterior, os consumidores estariam dispostos a salvar uma quantidade de dinheiro com o frete para produtos mais baratos do que produtos mais caros, semelhante ao que é indicado em Thaler (1980), pois o consumidor estará mais disposto a salvar $5 no frete em um produto que custa $25 do que um produto que custa $125.

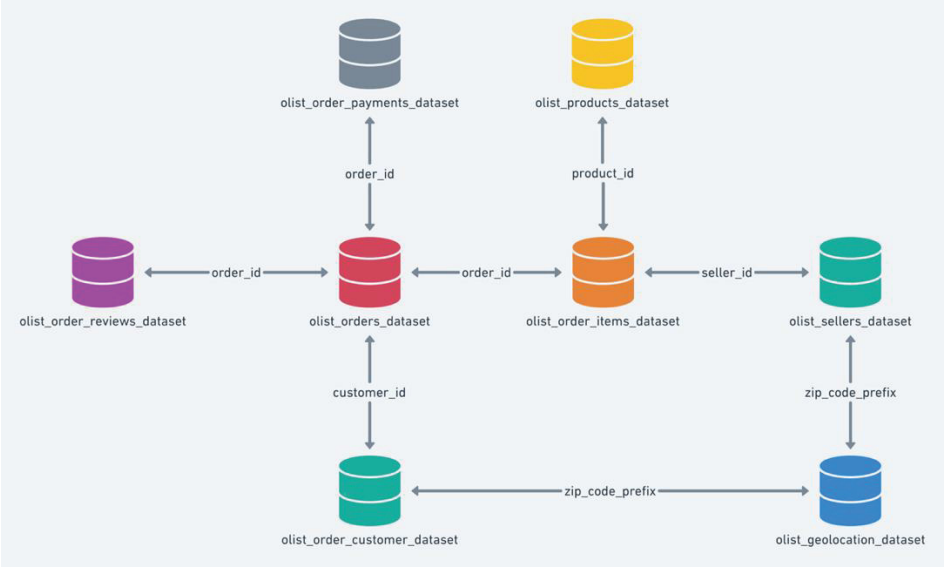
2.3. METODOLOGIA

2.3.1. BASE DE DADOS

Foram utilizados os dados de uma empresa de *E-commerce* brasileira chamada *Olist,* sendo que as bases de dados estão disponíveis no Kaggle. Primeiramente, a base de dados contendo os itens comprados (*olist_order_items_dataset*) foi sumarizada com o intuito de retirar pedidos sequenciais, isto é, casos em que a pessoa comprou mais unidades do mesmo produto, ficando na base somente o preço individual do produto (*price*), o preço do frete (freight_value) e a quantidade de itens comprados (*order_items_quantity*).

Em seguida, as bases de dados foram unidas seguindo a instrução contida no próprio Kaggle, como pode ser observada abaixo:

FIGURA 1: INTEGRAÇÃO DE BASE DE DADOS DA OLIST



Fonte: KAGGLE

A partir da base de dados de pedidos (*olist_orders_dataset*) foram trazidas as informações de *review* dos pedidos (*olist_order_reviews_dataset*), dados dos consumidores (*olist_order_customer_dataset)*, dados de itens comprados (*olist_order_items_dataset*), dados dos produtos (*olist_products_dataset*), dados dos vendedores (*olist_sellers_dataset)* e dados de geolocalização (*olist_geolocation_dataset)*.

A partir das latitudes e longitudes dos consumidores e dos vendedores foi calculada a distância do pedido (*distance*), utilizando o método Haversine[1]. Além disso, a partir do momento de compra (*order_purchase_timestamp*), foi criada uma coluna indicando o trimestre em que a compra foi realizada (*trim*), a hora da compra (*hour*), dia da compra (*day*) e dia da semana (*weekday*). Por fim, foi calculado o peso cúbico[2] (*cubic_weight*), dado pela seguinte fórmula:

$$peso\ cúbico = \frac{altura \times largura \times comprimento}{6000} \qquad (1)$$

Utilizando o peso cúbico (*cubic_weight*), foi calculado o peso utilizado no cálculo do frete (*weight_used*), onde caso o peso cúbico seja maior que 5 quilogramas é utilizado o peso cúbico, caso contrário é utilizado o peso do produto (*product_weight_g*). Por fim, foram filtrados apenas o *status* do pedido (*order_status*) que foram entregues e foram retirados da base os casos que estavam com informações faltantes no peso utilizado no cálculo do frete, na categoria do produto (*product_category_name*), no tamanho do nome do produto (*product_name_lenght*), no tamanho da descrição do produto (*product_description_lenght*), na quantidade de fotos (*product_photos_qty*) e na distância.

Após este processo foram removidas as latitude do consumidor (*customer_lat*), longitude do consumidor (*customer_lng*), latitude do vendedor (*seller_lat*), longitude do vendedor (*seller_lng*), peso do produto (*product_weight_g*), comprimento do produto (*product_length_cm*), altura do produto (*product_height_cm*), largura do produto (*product_width_cm*) e *status* do pedido (*order_status*).

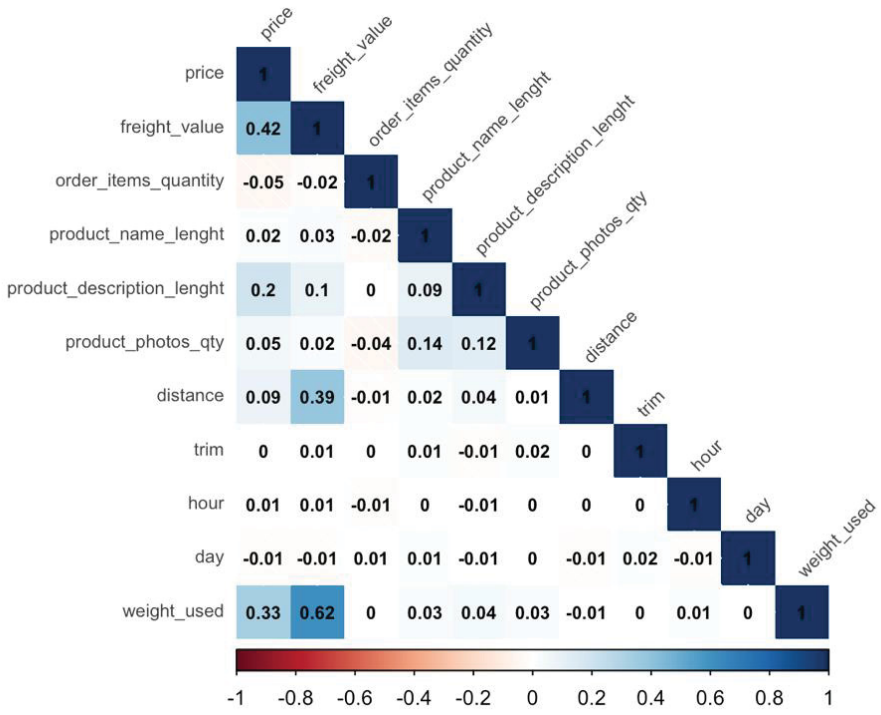
A partir da matriz de correlação mostrada abaixo, é possível observar que existe uma correlação de Pearson considerável de frete (*freight_value*) com o preço do produto (*price*), distância (*distance*) e peso utilizado (*weight_used*):

---

[1] Este método calcula a menor distancia em uma esfera entre dois pontos a partir de suas latitudes e longitudes

[2] A fórmula de peso cúbico é determinada pelo Correio, sendo que esta equação está disponibilizada no próprio site

GRÁFICO 1: CORRELAÇÃO DAS VARIÁVEIS ANTES DE AJUSTE DE FRETE



Fonte: Elaborado pelo autor

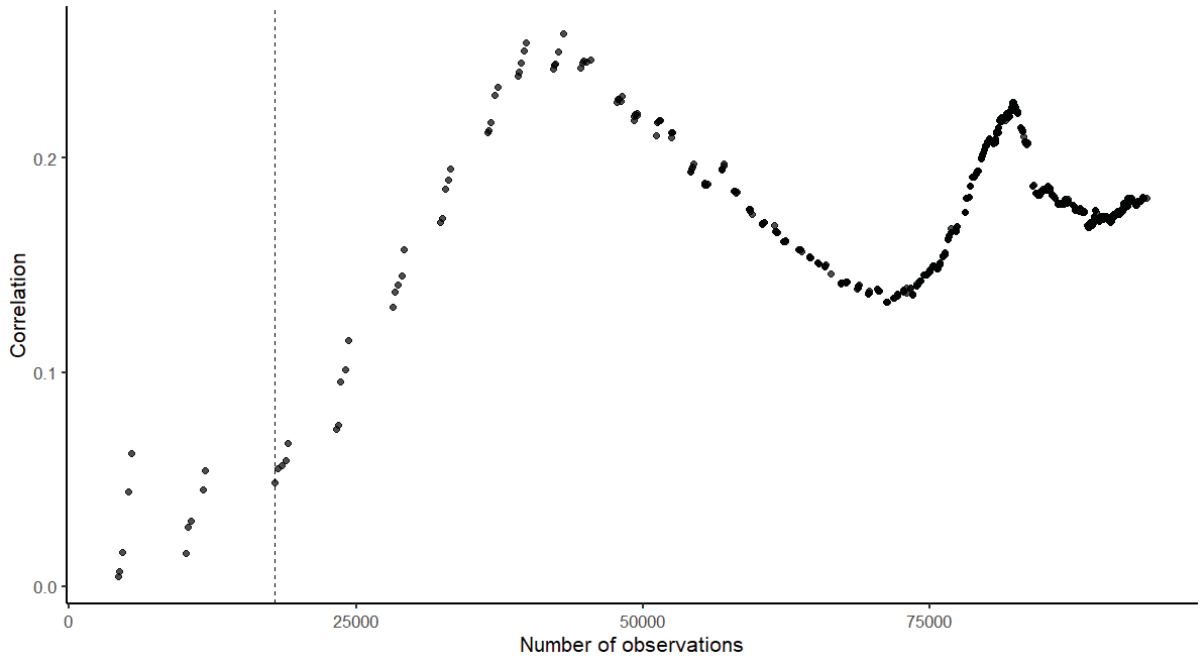
Esta alta correlação de frete com estas três variáveis é esperada, justamente porque o frete é calculado com base no peso do produto, na distância da entrega e um seguro facultativo de entrega do produto, o qual está em função do preço do produto. Desta maneira, a alta correlação de preço (*price*) com o frete (*freight_value*) se deve tanto ao seguro quanto à um problema de multicolinearidade entre peso utilizado (*weight_used*), preço do produto (*price*) e o frete (*freight_value*).

Para tirar o efeito do peso na relação de preço e frete, foi realizada uma análise de sensibilidade para filtrar observações com o peso ideal para minimizar a correlação entre as variáveis preço e peso do produto e maximizar o número de observações disponíveis na base de dados, sendo que o valor escolhido foi de peso menor ou igual a 200 gramas[3], dado que dentre os valores que minimizam a correlação, ele maximizar o número de observações de disponíveis, sendo este valor equivalente a 17.940 observações. O gráfico abaixo mostra a quantidade de

[3] O peso do produto de 200 gramas também é a moda do peso do produto presente na base de dados

observações disponíveis no eixo x e a correlação entre peso e preço para cada faixa de peso entre menor ou igual a 100 gramas e menor ou igual a 10 quilos, sendo que a linha tracejada mostra o ponto escolhido de 200 gramas:

GRÁFICO 2: SENSIBILIDADE DE CORRELAÇÃO ENTRE PESO E PREÇO DO PRODUTO



Fonte: Elaborado pelo autor

Nota-se pelo gráfico que pesos menores que 200 gramas diminuem significativamente o número de observações disponíveis para análise, e valores acima de 200 gramas aumentam a correlação para valores acima de 0.05 e aumentam o número de observações disponíveis, desta maneira, com intuito de minimizar a correlação, foi escolhido este valor de peso e utilizada 17940 observações.

Para retirar o efeito do preço no frete através do seguro, utilizou-se a tabela da ANTT (Associação Nacional de Transportes Terrestres) que indica qual alíquota de seguro deve ser utilizada por faixa de quilometragem. Para cada faixa de distância há uma alíquota específica a ser utilizada, como se pode observar na tabela abaixo:

TABELA 1: ALÍQUOTAS DE FRETE EM RELAÇÃO À DISTÂNCIA

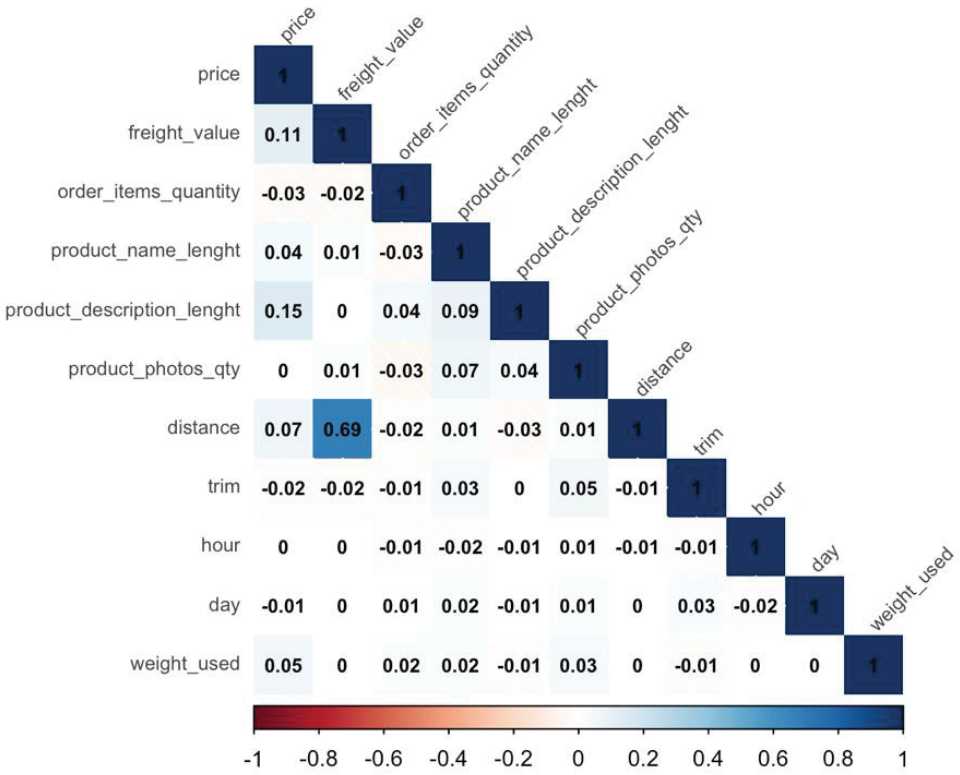
Alíquotas de frete-valor	
Distância (km)	Alíquota (%)
0000 - 0250	0,3
0251 – 0500	0,4
0501 - 1.000	0,6
1.001 – 1500	0,7
1.501 - 2.000	0,8
2.000 - 2.600	0,9
2.601 - 3.000	1,0
3.001 - 3.400	1,1
3.401 - 6.000	1,2
Coleta e entrega	0,15

Fonte: Associação Nacional de Transportes Terrestres (ANNT)

Assim, a partir da variável distância (*distance*) foi calculada a respectiva alíquota de cada pedido. Em seguida, foi calculado o valor do seguro através da multiplicação desta alíquota ao preço do produto e o valor do seguro foi subtraído do valor do frete (*freight_value*)[4]. Após realizar estes procedimentos, a correlação de Pearson entre preço e peso do produto diminuiu (evitando problemas relacionados à multicolinearidade) e a correlação de preço e com frete também diminuiu, como se pode observar na matriz de correlação abaixo:

[4] Nem todos os pedidos presentes na base de dados possuem necessariamente seguro facultativo, contudo não é possível diferenciar quais pedidos possuem este seguro e quais não possuem, desta maneira optou-se por subtrair de todos os casos para remover qualquer tipo de impacto do preço no frete que invalide o resultado

GRÁFICO 3: CORRELAÇÃO DAS VARIÁVEIS APÓS DE AJUSTE DE FRETE



Fonte: Elaborado pelo autor

Após o tratamento da base de dados, foram removidas as observações que não havia informações sobre o anúncio do produto (quantidade de fotos, quantidade de caracteres no título e na descrição), fazendo com que a base de dados possuísse 17670 observações e 14 variáveis. Das variáveis presentes na base de dados final, 3 variáveis eram categóricas: estado do consumidor (*customer_state*), categoria do produto (*product_category_name*) e dia da semana (*weekday*). As demais variáveis são numéricas, sendo que abaixo pode-se observar a estatística descritiva das mesmas:

TABELA 2: ESTATÍSTICA DESCRITIVA DAS VARIÁVEIS NUMÉRICAS

Variáveis numéricas	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
price	0,85	19,93	33,90	53,61	59,90	2749,00
freight_value	0,03	10,17	14,42	14,50	16,67	126,59
order_items_quantity	1,00	1,00	1,00	1,10	1,00	6,00
product_name_lenght	9,00	41,00	50,00	47,71	56,00	64,00
product_description_lenght	23,00	318,00	515,00	717,87	919,00	3985,00
product_photos_qty	1,00	1,00	1,00	2,11	3,00	13,00
distance	0,00	104,47	387,90	496,34	683,29	4155,42
trim	1,00	1,00	2,00	2,31	3,00	4,00
hour	0,00	11,00	15,00	14,67	19,00	23,00
day	1,00	8,00	16,00	15,73	23,00	31,00
weight_used	0,00	0,12	0,15	0,15	0,20	0,20

As variáveis categóricas foram transformadas em dummies e a base de dados foi dividida aleatoriamente em dois conjuntos, um para treinar os algoritmos (conjunto de treinamento) e um para testar o algoritmo (conjunto de teste). O primeiro consiste em 90% da base de dados (15906 observações) e 10% da base de dados no conjunto teste (1764 observações). A escolha da quebra de 90% para treinamento e 10% para teste se deve ao baixo número de observações disponíveis na base de dados (17670 observações), fazendo com que seja necessário separar uma parcela maior da amostra para treinamento do algoritmo.

2.3.2. HIPÓTESES

Hipótese 1: Os consumidores estão dispostos a pagar maiores níveis de frete conforme maiores são os preços do produto comprado (apontando para a presença de *Relative Thinking*);

Hipótese 2**:** O fenômeno de *Relative Thinking*, caso seja confirmado, afeta diferentemente os consumidores conforme a categoria de produto comprado, como sugerido por Hirshman et al. (2018);

2.3.3. ALGORITMOS UTILIZADOS

Para testar as duas hipóteses levantadas, será utilizada uma abordagem de Aprendizagem Supervisionada, mais especificamente uma abordagem de Regressão, onde a variável dependente é o frete (*freight_value*). As duas hipóteses buscam validar qual o impacto do preço do produto (*price*) no frete, além de compreender se este impacto varia entre os consumidores e entre tipos de produtos, desta maneira optou-se por utilizar quatro algoritmos distintos: Regressão Linear Múltipla, Árvore de Decisão e *Random Forest.*

James et al. (2013) apresenta critérios que diferenciam algoritmos de *Machine Learning*. O primeiro deles é se estes algoritmos são paramétricos ou não-paramétricos. Em geral, algoritmos paramétricos assumem explicitamente uma forma funcional de tal maneira que o modelo calcula apenas os parâmetros, contudo há o risco de que os pressupostos utilizados pela função assumida (por exemplo, linearidade) não ocorrem realmente nos dados analisados, fazendo com que a qualidade do modelo seja baixa. Os modelos não-paramétricos não assumem nenhuma forma funcional específica e, em geral, dependem de poucos pressupostos.

A segunda maneira de diferenciar estes algoritmos é baseado no *trade-off* de flexibilidade e acurácia destes modelos. Em geral, modelos com alta flexibilidade (geralmente, modelos não paramétricos) são difíceis de interpretar e modelos com baixa flexibilidade (geralmente, modelos paramétricos) possuem maior interpretabilidade.

Por fim, a terceira diferenciação que pode ser utilizada é a motivação pela qual o modelo será utilizado: inferência ou predição. No caso da inferência, o objetivo é compreender como Y está sendo afetado pelas variáveis dependentes, desta maneira é importante que o modelo seja interpretável. No caso da predição, o objetivo é apenas prever corretamente Y, desta maneira tendem a ser utilizados algoritmos com alta flexibilidade e baixa interpretabilidade.

Os algoritmos escolhidos no presente trabalho são baseados nestes três critérios. O primeiro deles é a Regressão Linear Múltipla (OLS), o qual é paramétrico, possui alta interpretabilidade e é apropriado para fazer inferência estatística (James et al., 2013). Desta maneira, este primeiro modelo permite entender as relações entre as variáveis de maneira simples através dos parâmetros estimados, contudo ele assume uma forma funcional linear e diversos pressupostos,

fazendo com que a acurácia do modelo seja baixa caso os pressupostos não sejam respeitados pelos dados utilizados.

O segundo algoritmo escolhido foi a Árvore de Decisão, o qual é não-paramétrico e é interpretável a partir da árvore de decisão gerada pelo modelo. A escolha deste algoritmo se deve a ele ser eficiente para lidar com comportamentos do consumidor (Garver, 2002), ao fato de seu resultado ser de fácil compreensão (Bounsaythip e Runsala, 2001), lida facilmente com variáveis discretas e contínuas na base de dados, seleciona automaticamente as variáveis relevantes, é robusto em relação a *outliers*, eficiente para lidar com grandes bases de dados (Murphy, 2012) e ,por fim, é amplamente utilizado para compreender o comportamento do consumidor (Ngai, Xiu e Chau, 2009). O grande problema com este algoritmo é que ele é instável, isto é, pequenas alterações no *input* podem mudar completamente a estrutura da árvore (Robert, 2014).

O terceiro algoritmo utilizado foi o *Random Forest*, o qual é não paramétrico e possui baixa interpretabilidade. O resultado deste algoritmo geralmente possui uma boa acurácia de predição, contudo seu resultado tende a ser difícil de compreender dado que ele consiste em diversas árvores de decisão e o resultado predito pelo algoritmo é o resultado agregado de todas estas árvores de decisão (Murphy, 2012), desta maneira é um modelo apropriado para questões preditivas. Uma das vantagens deste algoritmo é que ele possibilita diminuir a variância e a instabilidade do algoritmo da árvore de decisão (Murphy, 2012).

Desta maneira, os três algoritmos se complementam dado que a Regressão Linear Múltipla é altamente interpretável e pouco flexível (dado que é paramétrica). A Árvore de Decisão possui uma interpretabilidade intermediária (dado que não basta compreender o impacto da variável dependente através de parâmetros estimados) e maior flexibilidade, visto que é um modelo não paramétrico. Por fim, o algoritmo *Random Forest* é extremamente flexível mas possui baixa interpretabilidade, sendo apropriado para questões relacionadas à predição.

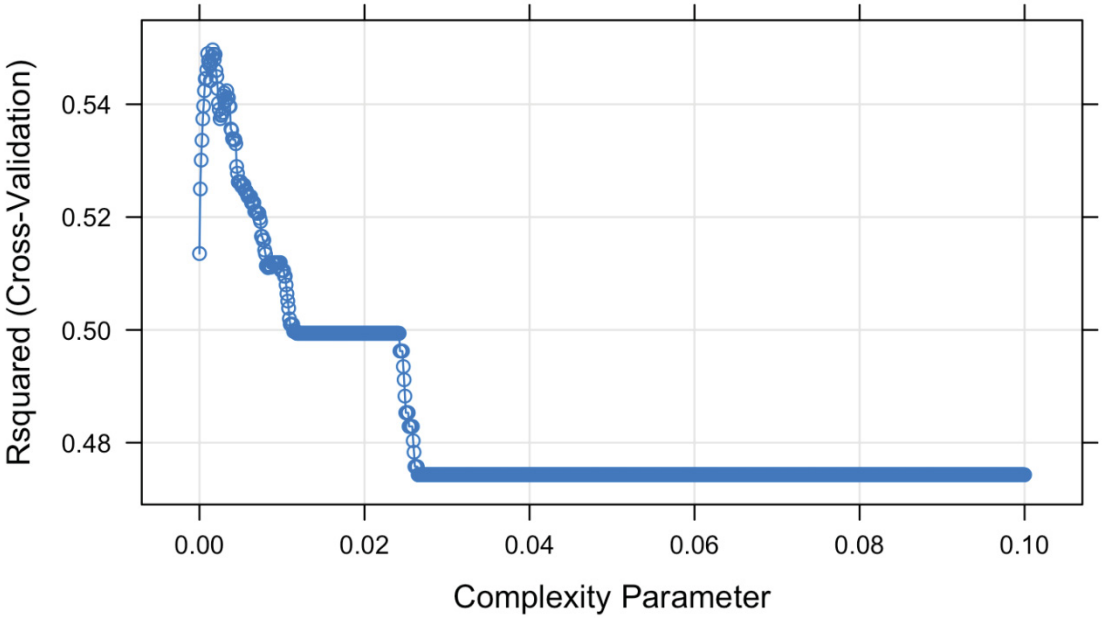
2.3.4. APLICAÇÃO DOS ALGORITMOS

Foi utilizado o software R para aplicação dos algoritmos utilizados, mais especificamente o pacote *Caret.* Nos casos de algoritmos não paramétricos (Árvore de Decisão e Random Forest) foi realizado técnicas de *Hyperparameter*

Tuning para selecionar os hiperparâmetros destes modelos que maximizar o R-quadrado.

Desta maneira, na Árvore de Decisão o hiperparâmetro relacionado à complexidade[5] foi otimizado com intuito de maximizar o R-Quadrado, sendo que o valor de Complexidade que otimiza o R-quadrado para este algoritmo é de 0.0015, conforme mostra o gráfico abaixo:

GRÁFICO 4: OTIMIZAÇÃO DE HIPERPARÂMETRO DA ÁRVORE DE DECISÃO



Fonte: Elaborado pelo autor

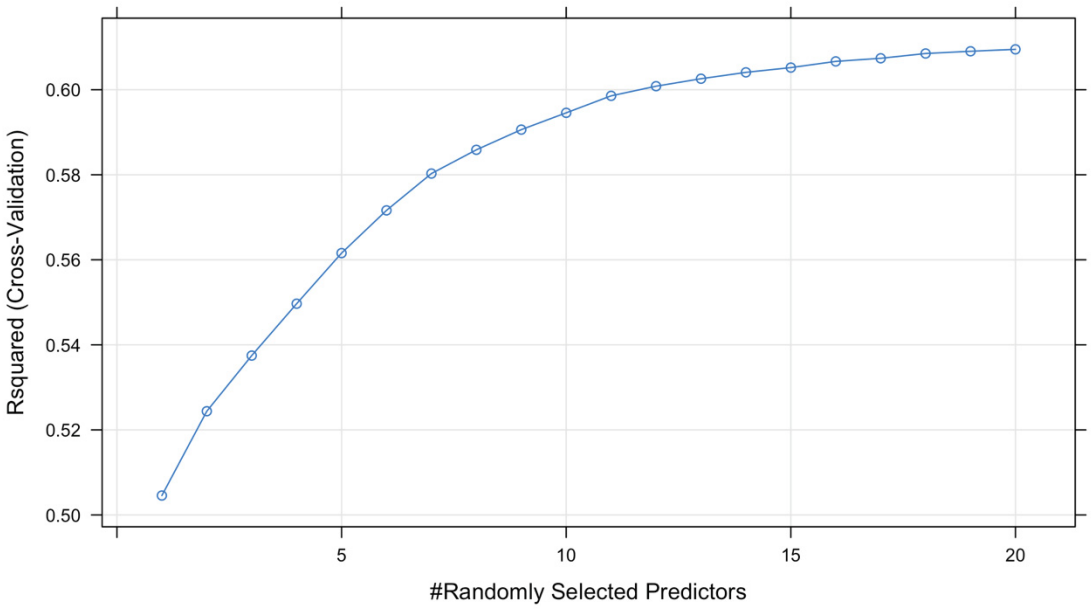
Quanto ao algoritmo de *Random Forest*, o hiperparâmetro otimizado foi a Número de variáveis para as quais é possível dividir cada nó[6], sendo que os valores possíveis para este hiperparâmetro vão de um até o número de variáveis presentes na base de dados. O valor deste hiperparâmetro que maximiza o R-quadrado do modelo é 18 e o resultado pode observado no gráfico abaixo:

---

[5] Foi utilizada a função *rpart* do pacote *rpart* disponível para o software R, sendo que o nome do hiperparâmetro é *cp*

[6] Foi utilizada a função *ranger* do pacote *ranger* disponível para o *software* R, sendo que o nome do hiperparâmetro é *mtry*

GRÁFICO 5: OTIMIZAÇÃO DE HIPERPARÂMETRO DE RANDOM FOREST



Fonte: Elaborado pelo autor

Além disso, para evitar problemas com *overfitting* dos três algoritmos, isto é, o algoritmo se adaptar exatamente com a estrutura dos dados do conjunto de treinamento de tal maneira que as predições para o conjunto teste não seja confiável, foi utilizado a técnica de *K-fold Cross Validation* onde a base de dados é dividida randomicamente em *K* grupos iguais, sedo que é selecionado um destes K grupos para ser o conjunto de validação e os demais K-1 são utilizados para treinamento do algoritmo, isso é repetido utilizando cada um dos K grupos como conjunto de validação, para observar se existem diferenças significativas na estrutura dos algoritmos e em sua performance (cuja métrica utilizada foi o R-Quadrado) entre cada um dos K grupos (Alpaydin, 2010). Neste trabalho o *K-fold Cross Validation* foi aplicado no conjunto de treinamento ao treinar os algoritmos e foi utilizado K = 10.

2.3.5. INTERPRETAÇÃO DOS RESULTADOS

Como descrito por James et al. (2013), o resultado da Regressão OLS é altamente interpretável, sendo que a interpretação se dá justamente pelos parâmetros calculados pelo modelo. Quanto à Árvore de Decisão, apesar de ser não-paramétrica, é possível interpretar a própria árvore gerada pelo modelo e

compreender quais foram as quebras que minimizaram o nível de impureza nos nós da árvore, fazendo com que seja um algoritmo de alta interpretabilidade (Bounsaythip e Runsala, 2001).

Contudo, devido ao *trade-off* entre interpretabilidade e a acurácia, apesar da Regressão OLS e Árvore de Decisão serem facilmente interpretáveis, elas tendem a ser menos acuradas. Desta maneira, utilizou-se o algoritmo de *Random Forest,* o qual tende a ser mais acurado, porém menos interpretável. Assim, para interpretar os resultados gerados pelo algoritmo de *Random Forest* foram utilizadas técnicas específicas para compreender quais as variáveis mais importantes para predição, interação entre variáveis na previsão do valor do frete e o impacto positivo ou negativo de cada variável independente do modelo.

A importância das variáveis para a predição do algoritmo foi calculada com base em Fisher, Rudin e Dominici (2018), sendo que metodologia se baseia em calcular primeiramente o erro do modelo original

$$e^{orig} = L\ (y,\ f\ (X)) \qquad (2)$$

Onde y é o valor real do frete e f(x) é o valor predito pelo modelo. Em seguida, é gerada uma matriz de permutação (x^{perm}) para cada uma das variáveis e é calculado o erro do modelo baseado nas previsões realizadas pelo algoritmo com os dados permutados:

$$e^{perm} = L\ (Y,\ f\ (X^{perm})) \qquad (3)$$

Por fim, é calculada a importância para cada variável j através da seguinte fórmula:

$$FIj = e^{perm} / e^{orig} \qquad (4)$$

A interação entre as variáveis foi calculada com base em Friedman e Popescu (2008), cujo o teste é denominado de "*H-statistic".* O teste é dado pela comparação da soma das funções de dependência parcial das variáveis *j* e *k* ($PD_{jk}(x_j, x_k)$) com a decomposição das funções de dependência parcial da variável *j* ($PD_j(x_j)$) e da

variável *k* ($PD_k(x_k)$) sendo que o teste é feito para cada observação *i* da base de dados, como pode ser visto na equação abaixo:

$$H_{jk}^2 = \sum_{i=1}^{n} \left[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right]^2 / \sum_{i=1}^{n} PD_{jk}^2(x_j^{(i)}, x_k^{(i)}) \quad (4)$$

O resultado do teste vai de 0 à 1, sendo que quando o resultado é 0 não há interações entre as variáveis e quando é 1 todo o impacto destas variáveis se devem à interação somente.

Para compreender o impacto positivo ou negativo de cada variável na predição foi utilizado *Individual Conditional Expectation* (ICE) baseado em Goldstein et al. (2017). Esta metodologia foi desenvolvida a partir de *Partial Dependence Plot* (PDP) de Friedman (2001), a qual mostra o efeito marginal da variável independente no resultado previsto pelo algoritmo. Para determinada variável x_S, a função de dependência parcial é calculada considerando o impacto das demais variáveis (x_C) na predição do algoritmo, como pode ser observado na equação abaixo:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)\, d\mathbb{P}(x_C) \quad (5)$$

O PDP pode mostrar entre a variável dependente e a variável independente (x_S) possui uma relação linear ou mais complexa. O ICE mostra a dependência parcial da variável dependente e a variável independente para cada uma das observações da base de dados, sendo que o PDP nada mais é do que a média da dependência parcial de cada uma das observações. A vantagem de utilizar o ICE em detrimento do PDP se deve ao fato do segundo poder obscurecer relações heterogêneas entre as variáveis, as quais podem ocorrer devido à interação entre as variáveis.

Por fim, o último método utilizado foi *Shapley-values*, baseado no modelo de teoria dos jogos de coalizão de Shapley (1953), onde os jogadores recebem um retorno financeiro baseado em sua contribuição no retorno financeiro total, recebendo um “lucro” devido à sua cooperação. Na metodologia de *Shapley-values* as variáveis competem para prever o resultado, sendo que o valor para cada variável em cada observação nada mais é que a contribuição da variável para a predição de uma observação comparada com a média da predição na base de dados.

Para calcular este valor foi utilizada a aproximação de Štrumbelj e Kononenko (2014), onde para calcular o *shapley-value* da variável *j*, uma quantidade aleatória de valores das demais variáveis (diferentes de *j*) é substituída valores aleatórios de outras observações da base de dados. Após isto, o algoritmo aplicado para prever no conjunto considerando a variável j e os valores substituídos aleatoriamente. Em seguida o algoritmo é aplicado novamente sobre estes valores substituídos aleatoriamente, mas sem a variável j. A diferença entre da previsão com a variável *j* e sem a variável *j* é computada, sendo que isto é feito para M diferentes interações e o *shapley-value* é a média da diferença de cada uma das interações. A equação abaixo descreve esse processo:

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}\left(x_{+j}^{m}\right) - \hat{f}\left(x_{-j}^{m}\right) \right) \qquad (6)$$

Desta maneira, o *shapley-value* mostra se o impacto da variável é positivo ou negativo, considerando o impacto individual de cada uma das outras variáveis, para cada uma das observações, possibilitando entender como cada variável contribuiu para a predição realizada pelo algoritmo.

A partir destas técnicas de interpretação de algoritmos é possível testar as hipóteses levantadas utilizando o algoritmo de *Random Forest.*

2.4. RESULTADOS

Os resultados encontrados serão discutidos em três tópicos, o primeiro deles discutindo os resultados obtidos com a Regressão Linear Múltipla, o segundo discutindo os resultados da Árvore de Decisão e o terceiro se refere ao algoritmo *Random Forest*. Em cada um dos tópicos, os resultados encontrados serão analisados baseado nas duas hipóteses levantadas, as quais são:

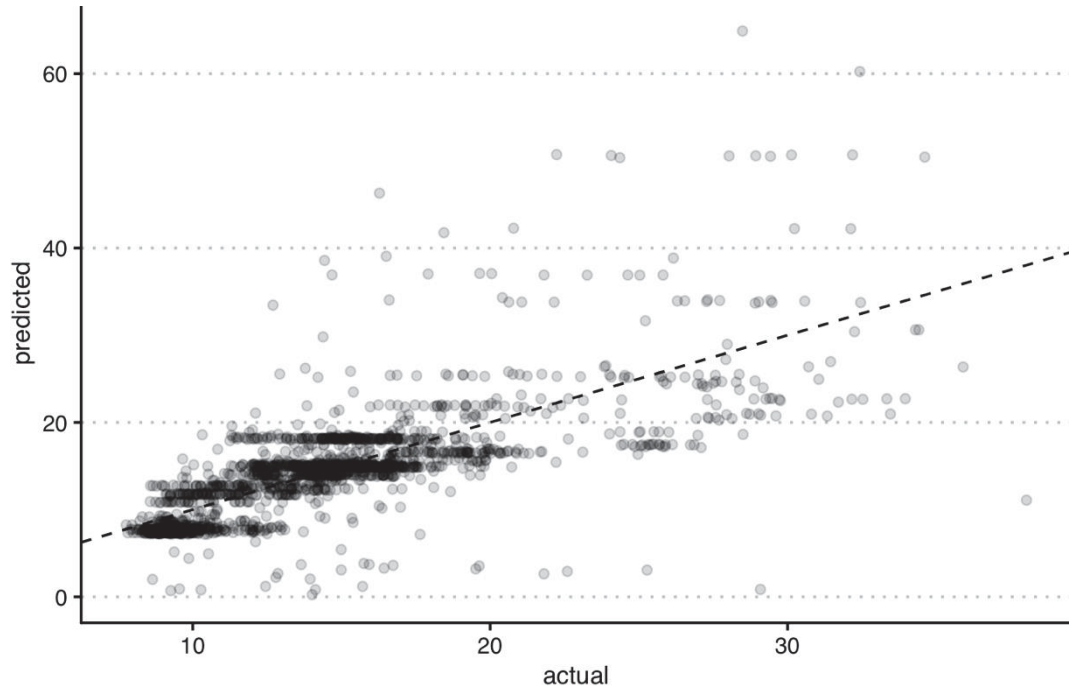
- Hipótese 1: Os consumidores estão dispostos a pagar maiores níveis de frete conforme maiores são os preços do produto comprado (apontando para a presença de *Relative Thinking*);

- Hipótese 2**:** O fenômeno de *Relative Thinking*, caso seja confirmado, afeta diferentemente os consumidores conforme a categoria de produto comprado, como sugerido por Hirshman et al. (2018);

2.4.1. REGRESSÃO LINEAR MÚLTIPLA

A grande vantagem da utilização do modelo de Regressão Linear Múltipla no presente trabalho é sua alta interpretabilidade, permitindo compreender não apenas a magnitude do impacto das variáveis independentes, como também se este impacto é positivo ou negativo, todavia o algoritmo tende a ter menor acurácia que os demais. O gráfico abaixo mostra a relação do que foi predito pelo algoritmo[7] em comparação com o valor real no conjunto teste:

GRÁFICO 6: COMPARAÇÃO DE VALORES PREDITOS PELA REGRESSÃO LINEAR MÚLTIPLA E VALORES DE FRETE REAIS



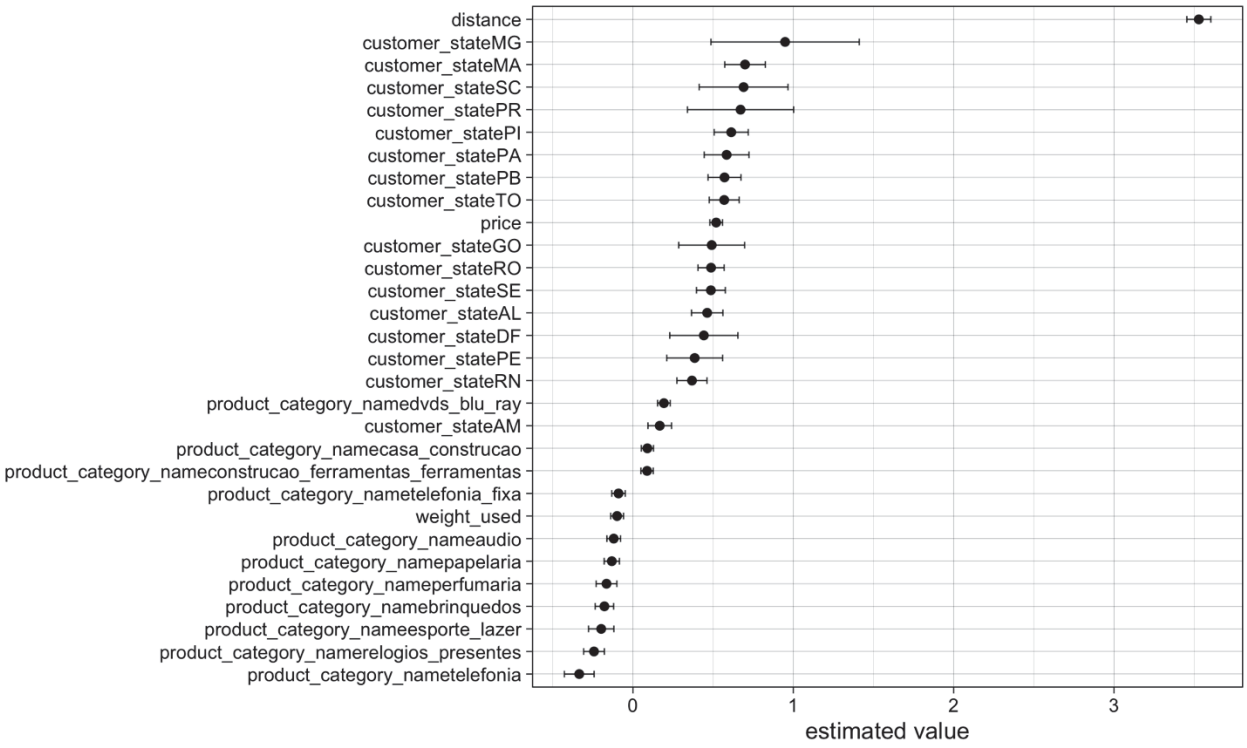
Fonte: Elaborado pelo autor

[7] Foram realizados testes dos seguintes pressupostos da Regressão Linear Múltipla: média dos resíduos igual a zero, homocedasticidade dos resíduos, autocorrelação dos resíduos, existência de valores influentes (medidos pela distância de Cook) e multicolinearidade. Os resultados se encontram em anexo 2.

Observa-se que a grande parte dos pontos estão acima da reta tracejada, indicando que o valor previsto pelo algoritmo é semelhante ao valor real, sendo que R-quadrado do modelo no conjunto de treinamento foi de 0.529 e o R-quadrado ajustado foi de 0.5255, indicando que a quantidade de variáveis não está afetando negativamente o modelo. Além disso, no conjunto de teste o R-quadrado foi de 0.541 e o R-quadrado ajustado foi de 0.5367, mostrando que o modelo já treinado generalizou bem para o conjunto de teste.

A tabela com os resultados dos modelos se encontra no Anexo A, sendo que o gráfico abaixo mostra o valor estimado e o desvio padrão das variáveis que foram significativas até 10%, com exceção do intercepto:

GRÁFICO 7: BETAS ESTIMADOS E DESVIO PADRÃO DAS VARIÁVEIS SIGNIFICATIVAS (p < 0.10) DA REGRESSÃO LINEAR MÚLTIPLA

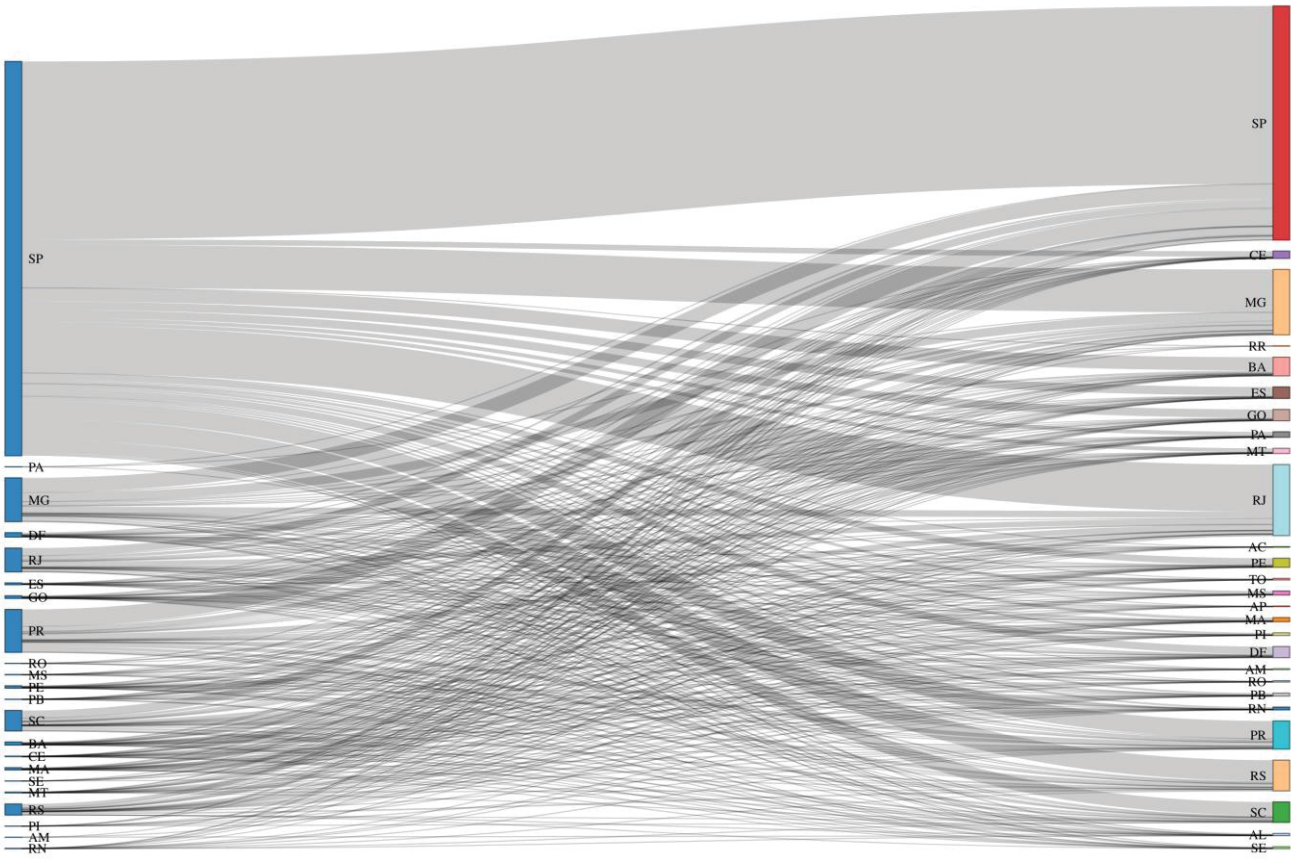


Fonte: Elaborado pelo autor

Nota-se que a distância (*distance*) é a variável que possui maior impacto no frete, com um beta estimado de 3,529. Este valor é esperado, dado que após ser retirado o impacto do peso do produto e do preço no frete (por meio do seguro e do peso do produto), a única variável que afeta diretamente o frete é a distância, fazendo com que seu impacto seja o maior dentre todas as variáveis.

Além disso, o frete é sensível a variáveis relacionadas ao estado do consumidor, sendo que consumidores de Minas Gerais (MG), Santa Catarina (SC), Maranhão (MA), Piauí (PI), Paraíba (PB), Tocantins (TO), Pará (PA), Paraná (PR) e outros tendem a pagar um valor de maior. Isso ocorre porque o estado do consumidor está intimamente relacionado com a distância, dado que a maior parte dos vendedores se encontram em São Paulo, fazendo com que os estados mais longes paguem um valor de frete maior dado que a distância em relação à São Paulo é maior. O gráfico abaixo mostra o fluxo de mercadorias, mostrando que o maior ofertante e demandante é São Paulo, seguido pelos demais estados do Centro-Sul do Brasil:

GRÁFICO 8: FLUXO DE ORIGEM DOS PRODUTOS A PARTIR DO ESTADO DO VENDEDOR PARA ESTADO DO CONSUMIDOR



Fonte: Elaborado pelo autor

O preço se mostrou uma variável com um impacto relativamente alto no frete, quando comparado às demais variáveis, sendo que o valor estimado foi de 0.519 e foi significativo à 0,1%. Isto indica que, mesmo retirando os impactos do preço no frete por meio do seguro e por meio do peso do produto, continua havendo um impacto positivo e estatisticamente significativo do preço no valor do frete, no qual preços maiores fazer com que o valor de frete pago tenda a ser maior.

Este resultado vai a favor da Hipótese 1, indicando que ao comprar produtos com maiores preços, os consumidores estão pagando maiores valores de frete. Este é justamente o resultado esperado a partir do que foi observado nos trabalhos de Thaler (1980), Kahneman e Tversky (1981) e Azar (2007), dado que os consumidores não olham o frete como um valor absoluto, mas como um valor relativo ao preço do produto, fazendo com que os fretes maiores pareçam mais atrativos quando os preços do produtos são maiores. Desta maneira, assim como consumidores estão mais dispostos a poupar $5 em um rádio de $25 do que em uma televisão de $500 (Thaler, 1980), o resultado indica que os consumidores estão menos dispostos a poupar no valor do frete quando o preço do produto é maior.

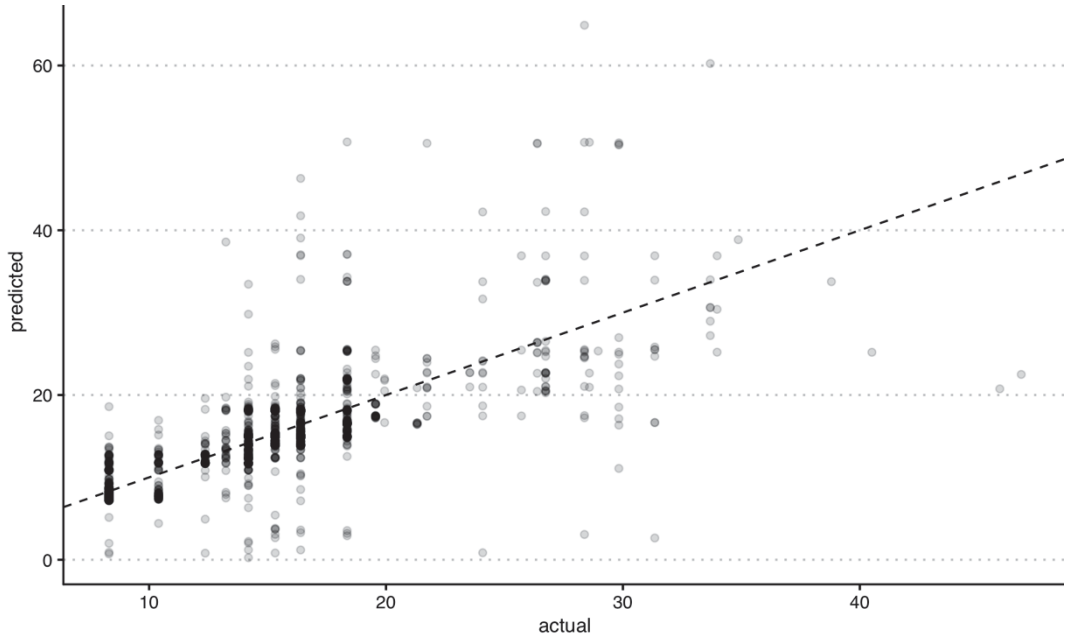
Por fim, observa-se que existe um impacto heterogêneo das categorias dos produtos no valor do frete. Por exemplo, o frete tende a ser maior em produtos relacionados às categorias “DVD Blu-Ray”, “Casa e Construção” e “Ferramentas”. Contudo ele tende a ser menor em produtos presentes nas categorias “Brinquedos”, “Esporte e Lazer”, “Telefonia” e outras. Isso indica que a categoria de produto afeta de maneira diversa o valor do frete, contudo isso não está relacionado ao fato de que determinadas categorias possuem produtos mais pesados que outros, dado que foram filtrados apenas produtos com pesos iguais ou menores que 200 gramas.

Isto apenas mostra que diferentes categorias de produtos impactam o valor do frete de diferentes maneiras, e isto não pode ser associado com a Hipótese 2, dado que o objetivo da mesma é observar se impacto do preço no frete muda por categoria de produto e não se o valor do frete pago varia conforme a categoria do produto. Para testar esta hipótese é necessário observar se há interação entre preço e categoria do produto para prever o valor do frete pago, sendo que isto será feito mais adiante.

2.4.2. ÁRVORE DE DECISÃO

Em contraste ao algoritmo de Regressão Linear Múltipla possui maior acurácia, dado que é não paramétrico e mais flexível, contudo possui menos interpreabilidade, apesar de ainda ser interpretável. Isto possibilita aumentar a acurácia do algoritmo e comparar o resultado gerado pela Árvore de Decisão com o resultado da Regressão Linear Múltipla. O gráfico abaixo mostra a relação entre o valor de frete predito pelo algoritmo e o valor real do frete:

GRÁFICO 9: COMPARAÇÃO DE VALORES PREDITOS PELA ÁRVORE DE DECISÃO E VALORES DE FRETE REAIS

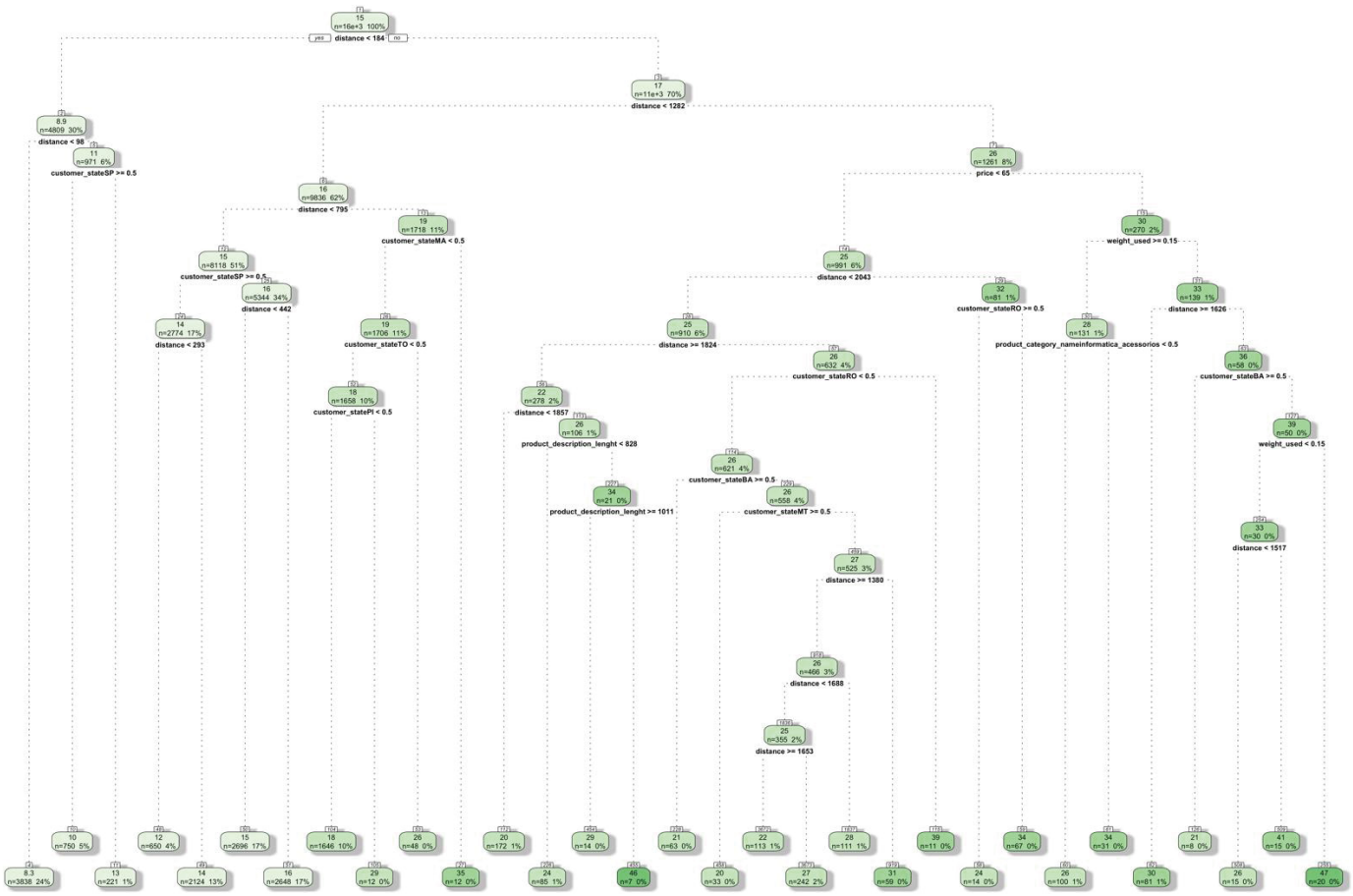


Fonte: Elaborado pelo autor

A performance do algoritmo de Árvore de Decisão foi superior ao algoritmo de Regressão Linear Múltipla, sendo que o R-quadrado do modelo foi de 0.5496511 e o RMSE (*Root Squared Mean Error*) foi de 4.559. No conjunto teste, o R-Quadrado foi de 0.5628824 e o RMSE foi de 4.298, indicando que não houve *overfitting* e que o algoritmo previu bem no conjunto teste.

A árvore de decisão gerada após o treinamento do algoritmo pode ser observada abaixo:

GRÁFICO 10: ÁRVORE DE DECISÃO ESTIMADA PARA EXPLICAR O VALOR DE FRETE PAGO PELO CONSUMIDIR



Fonte: Elaborado pelo autor

Dentro de cada um dos nós da Árvore de Decisão existem três informações: o valor de frete previsto pelo algoritmo para as observações que se encontram dentro do nó, a quantidade de observações e a porcentagem das observações da base de dados contida dentro do nó. Para variáveis numéricas a quebra é feita utilizando um limite (definido pelo algoritmo) no qual a variável é quebrada para melhor separar os valores de frete, por exemplo: a primeira quebra realizada pela árvore é baseada na distância, sendo que o limite escolhido pelo algoritmo foi o valor de 184. Em casos de variáveis *dummies*, a quebra é feita utilizando o valor de 0.5, sendo que se a variável for menor que 0.5, implica que a *dummie* é igual à zero, caso contrário ela será igual à 1.

A variável utilizada na primeira quebra e na segunda quebra da árvore de decisão é a distância, indicando que esta variável foi selecionada como a mais relevante para separar os maiores valores de frete e os menores valores de frete,

sendo que isso vai de acordo com o alto valor estimado para a variável distância na Regressão Linear Múltipla, indicando que esta é a variável que melhor explica o valor do frete.

Além disso, diversas quebras realizadas pela árvore são baseadas no estado do consumidor, indicando que esta variável desempenha um papel importante para prever o valor do frete pago. Isso se assemelha ao que foi encontrado na Regressão Linear Múltipla, na qual as *dummies* de estado do consumidor possuíam diferentes impactos no valor do frete e possuíam um valor estimado relativamente alto quando comparado às demais variáveis. Isso se deve justamente ao fato de a maior parte dos ofertantes serem do centro-sul do Brasil, fazendo com que estados brasileiros mais distantes dessa região tendam a pagar um valor de frete maior, visto que estão mais distantes dos ofertantes.

Por fim, observa-se que na terceira quebra, a variável escolhida pelo algoritmo foi o preço. Em casos que a distância é maior que 1282 quilômetros, se o preço for menor que 65 o valor previsto de frete R$25, caso contrário o valor previsto é R$30. Desta maneira, o algoritmo considerou nesta quebra que para preços maiores o valor do frete tende a ser maior, de maneira semelhante ao que foi encontrado pela Regressão Linear Múltipla, apontando novamente em direção da Hipótese 1 e ao que foi encontrado por Thaler (1980), Kahneman e Tversky (1981) e Azar (2007).

Após a árvore realizar a quebra baseada no preço (preços maiores que 65 e menores que 65), observa-se que, nos casos em que o preço é maior que 65 e em que o peso utilizado é maior ou igual a 0.15, a variável escolhida foi uma *dummie* de categoria de produto relacionada a acessórios de informática. De acordo com a árvore, caso o preço seja maior que 65 e a categoria do produto seja de acessórios de informática, o valor do frete tende a ser maior do que para outras categorias de produtos. Isso aponta justamente para o que é discutido na Hipótese 2, na qual o impacto do preço no valor do frete muda conforme a categoria do produto, visto que o fenômeno de *Relative Thinking* varia conforme a categoria do produto (Hirshman et al., 2018).

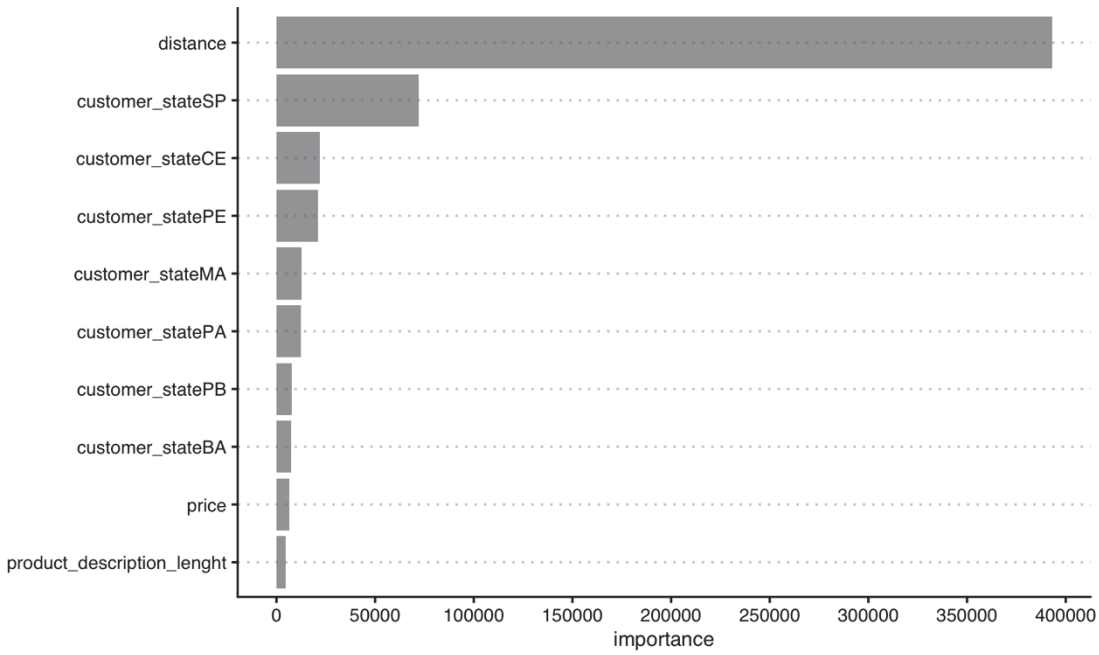
Um ponto relevante de discussão é o fato de o preço não ser uma variável recorrentemente escolhida pelo algoritmo para dividir os grupos, dado que ele apareceu apenas uma vez na árvore de decisão gerada. Isso não necessariamente implica que o preço não está exercendo impacto algum nas

quebras em que ele não aparece, apenas implica que outras variáveis separam melhor os grupos baseados no valor de frete pago. Desta maneira, pelo fato de a distância separar melhor as observações dentro do nó da árvore, ela aparece de maneira mais recorrente que o preço, sendo que isso é esperado dado o impacto direto da distância no valor do frete.

É possível notar que apesar da acurácia maior da Árvore de Decisão em relação à Regressão Linear Múltipla, sua interpretabilidade é menor, principalmente por causa do tamanho da árvore. Desta maneira, para facilitar a compreensão do impacto das variáveis na variável dependente, podemos calcular a importância das variáveis na previsão do frete, sendo que esta metodologia é baseada Atkison e Therneau (2010), na qual o critério utilizado para definir importância da variável é quão bem a mesma separou os grupos analisados.

O gráfico abaixo mostra as 10 variáveis com maior importância na previsão do frete, calculado a partir da metodologia de Atkison e Therneau (2010):

GRÁFICO 11: IMPORTÂNCIA DAS VARIÁVEIS NA ÁRVORE DE DECISÃO



Fonte: Elaborado pelo autor

É possível notar que a variável com maior importância é a distância, seguido por variáveis relacionadas ao estado do consumidor e em nono lugar observa-se o preço como a variável mais importante. Isto evidencia o que já foi

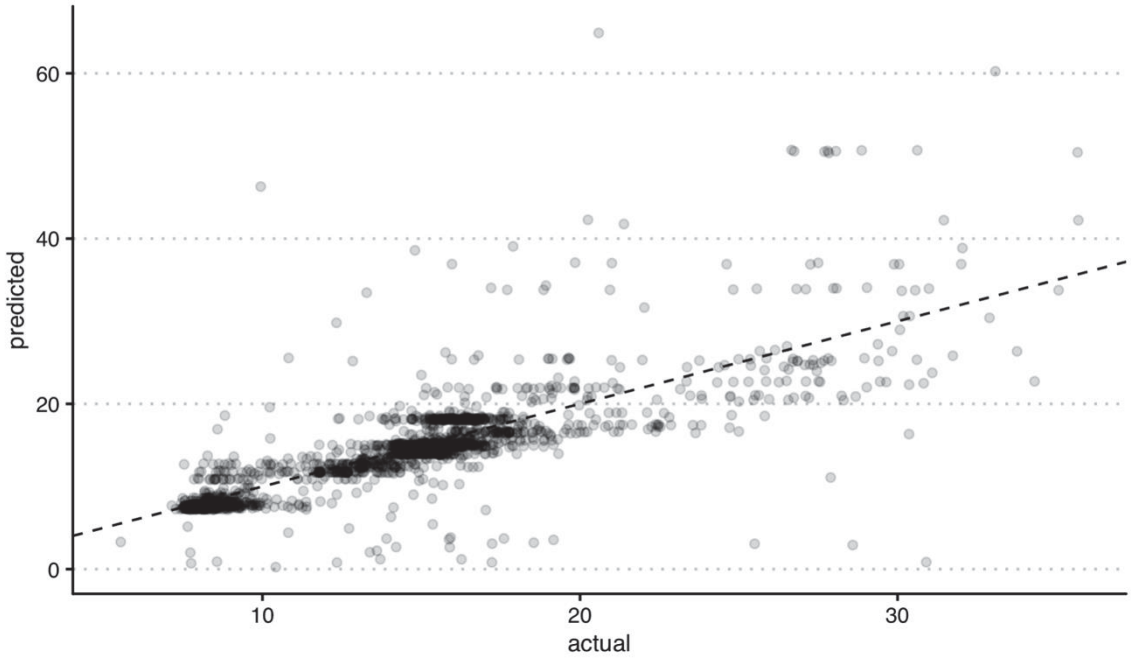
discutido anteriormente, onde é natural a distância possuir um impacto extremamente alto, dado seu impacto direto no valor do frete. Além disso, o impacto do estado do consumidor está associado à distância, visto que maior parte dos vendedores se encontram em São Paulo e no centro-sul do Brasil, fazendo com que a *dummie* relacionada aos consumidores do estado de São Paulo e de estados distantes de São Paulo (Ceará, Pernambuco etc) sejam importante para prever o valor do frete. Por fim, observa-se que o preço possui um impacto relevante no valor do frete (considerando os impactos altos e justificáveis da distância e estado do consumidor), além de ter sido observado na árvore que este impacto é positivo, isto é, o valor de frete pago tende a ser maior conforme maior o preço do produto, conforme descrito na Hipótese 2.

2.4.3. RANDOM FOREST

Como já discutido anteriormente, o algoritmo *Random Forest* possui uma interpretabilidade extremamente baixa quando comparada aos demais algoritmos aplicados, contudo tende a ter uma performance muito maior. Devido a isto, serão utilizadas técnicas de interpretação de algoritmos de *Machine Learning* para trazer maior interpretabilidade aos resultados obtidos por meio deste algoritmo, conforme discutido na seção 3.5.

O gráfico abaixo mostra a relação entre o que foi predito pelo algoritmo e o valor real do frete pago pelo consumidor:

GRÁFICO 12: COMPARAÇÃO DE VALORES PREDITOS POR R*ANDOM FOREST* E VALORES DE FRETE REAIS

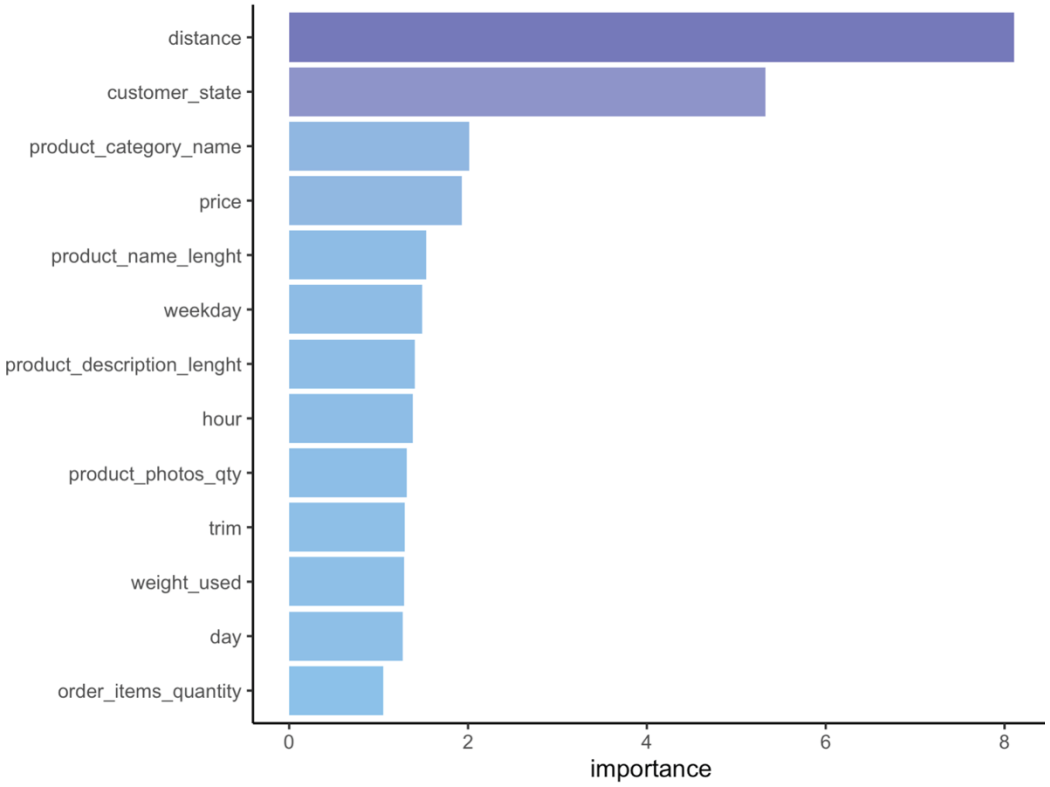


Fonte: Elaborado pelo autor

A acurácia do modelo foi maior em relação à Regressão Linear Múltipla e à Árvore de Decisão. O r-quadrado no conjunto de treinamento foi 0.61 e o RMSE foi de 4.25. No conjunto teste o r-quadrado foi de 0.62 e o RMSE foi de 3.97, indicando que não houve *overfitting* do modelo e a generalização foi boa para o conjunto de dados referente ao teste.

A primeira técnica para interpretar o algoritmo é a importância das variáveis, baseado em Fisher, Rudin e Dominici (2018). Através dela podemos saber a importância de cada variável no modelo para prever o valor do frete, o gráfico abaixo mostra a importância de cada uma das variáveis:

GRÁFICO 13: IMPORTÂNCIA DAS VARIÁVEIS NA PREDIÇÃO DE FRETE PARA *RANDOM FOREST*



Fonte: Elaborado pelo autor

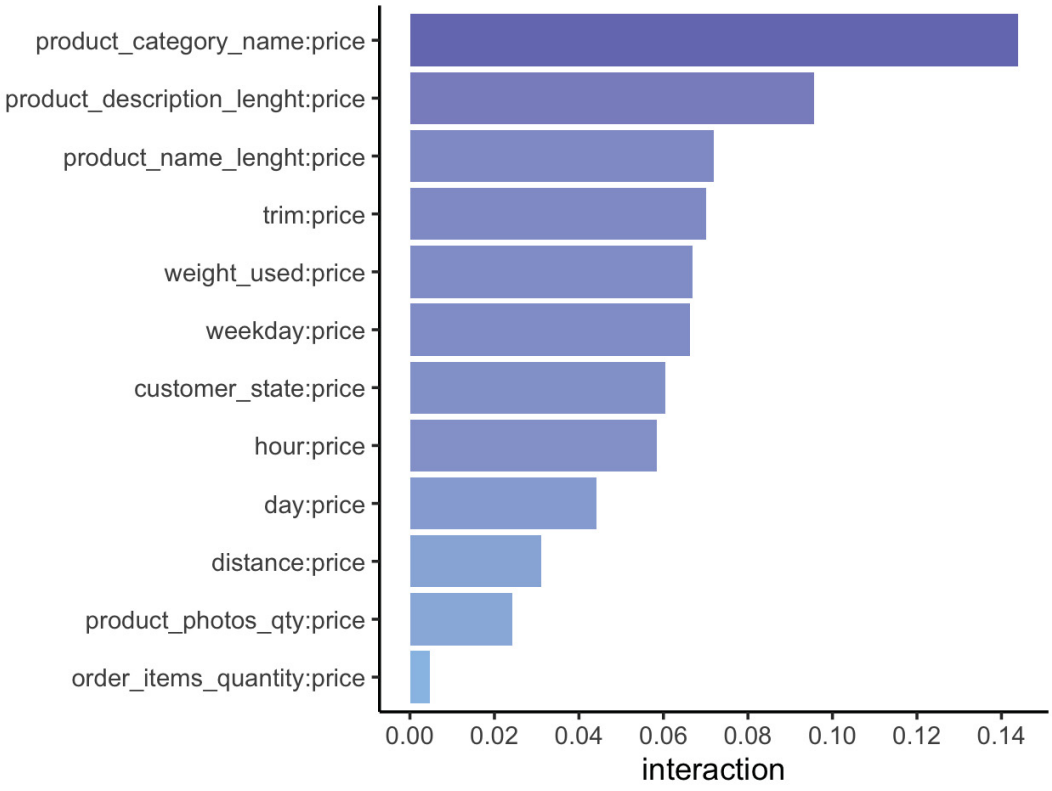
Os valores relativos à importância presentes no eixo X indicam quanto aumentou o erro do modelo (medido pelo RMSE) ao permutar o valor da variável. Semelhante ao que foi observado nos dois modelos anteriores, nota-se que a Distância possuía a maior importância preditiva, dado que o erro aumenta em mais de 8 vezes ao permutar esta variável, seguido pelo Estado do Consumidor. Conforme discutido no resultado da Regressão Linear Múltipla e na Árvore de Decisão, é esperado que estas duas variáveis possuam a maior importância.

A terceira variável com maior importância para prever o valor do frete é a Categoria do Produto. Este resultado vai de acordo com o que foi encontrado na Regressão Linear Múltipla, dado que para diferentes categorias de produtos foi observado diferentes impactos no valor do frete, contudo ainda não é possível afirmar que este resultado aponta para a hipótese 2, dado que para isso é necessário que haja interação entre preço e categoria do produto no valor de frete pago pelo consumidor, fazendo com que o impacto do preço no valor do frete varie por categoria de produto.

Por fim, a quarta variável mais relevante para prever o valor de frete pago pelo consumidor foi o preço, sendo que o erro do modelo é duas vezes maior quando os valores dessa variável são alterados. Isso indica que o preço, mesmo retirando seu impacto do valor do frete (através do peso do produto e do seguro), possui um impacto relevante no valor do frete pago pelo consumidor, contudo não se pode afirmar com este resultado se o impacto é positivo ou negativo.

Para testar a hipótese 2 é necessário compreender se há interação entre a Categoria do Produto e o Preço do mesmo para prever o valor de frete pago pelo consumidor, caso haja interação, isto indica que o impacto do preço no frete muda conforme a categoria do produto, apontando para o que foi encontrado por Hirshman et al. (2018), onde o efeito do fenômeno de *Relative Thinking* varia conforme o tipo de produto. Para calcular a interação entre variáveis foi utilizada a metodologia proposta por Friedman e Popescu (2008), onde foi calculada a interação do Preço com cada uma das outras variáveis, sendo que o resultado pode ser observado no gráfico abaixo:

GRÁFICO 14: INTERAÇÃO ENTRE VARIÁVEIS COM PREÇO NA PREDIÇÃO DE FRETE USANDO RANDOM FOREST



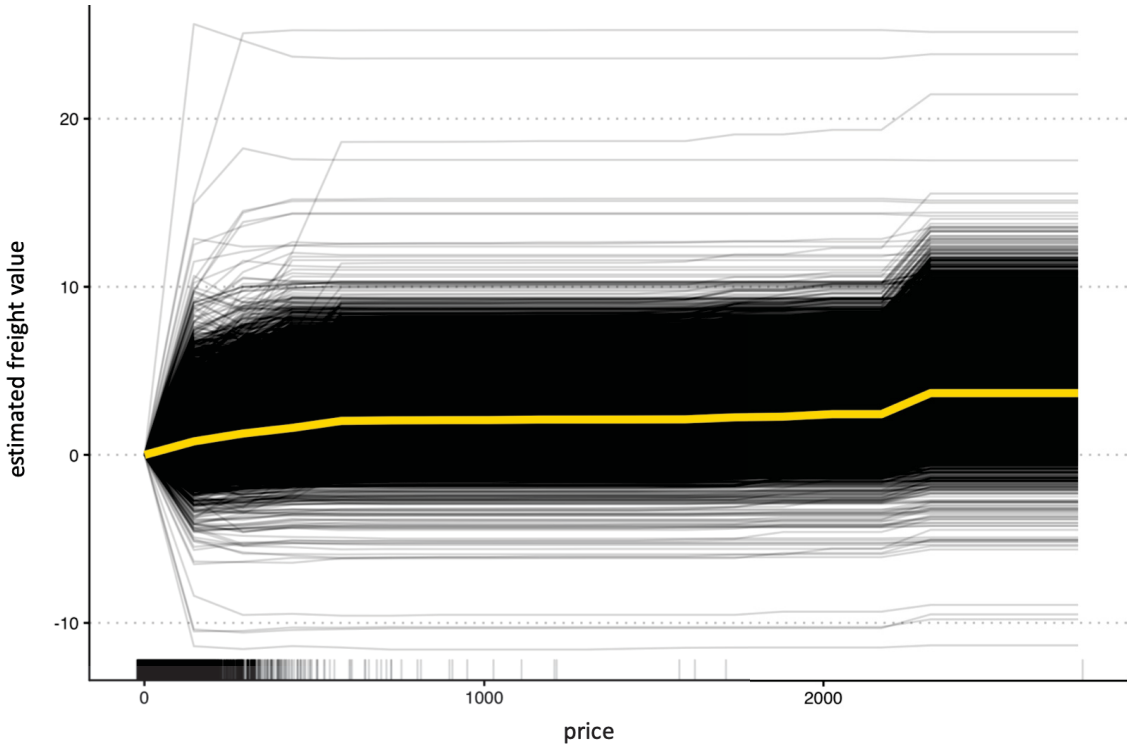
Fonte: Elaborado pelo autor

O valor calculado para a interação entre variáveis vai de 0 a 1, sendo que 0 indica que as duas variáveis praticamente não interagem para prever valor do frete e 1 significa que todo o impacto das variáveis no valor do frete se deve à interação (e não às variáveis individualmente). Observa-se que a variável que mais interage com o Preço para prever o valor do frete é a Categoria do Produto, sendo que isto aponta diretamente para a Hipótese 2, na qual o fenômeno de *Relative Thinking* afeta diferentemente os consumidores conforme a categoria de produto comprado.

Conforme encontrado por Hirshman et al. (2018) a partir dos resultados obtidos Hasting e Shapiro (2013), foi observado através de experimentos que o fenômeno de *Relative Thinking* se mostrou presente em 5 tipos de produtos distintos com exceção da gasolina, mostrando que o impacto deste fenômeno cognitivo varia de acordo com o tipo de produto. O resultado encontrado pela interação entre variáveis aponta para a mesma direção, sendo que o impacto do preço no valor do frete depende da categoria do produto.

Foi observado que o preço exerce uma importância relativamente alta na predição do valor do frete pago através do método de em Fisher, Rudin e Dominici (2018), contudo é necessário compreender qual a relação entre preço e frete. Espera-se que a relação entre as duas variáveis seja positiva, sendo este um indicativo da presença de *Relative Thinking* (conforme discutido na Hipótese 1). Para compreender a relação entre o preço e o frete foi utilizada a metodologia de *Individual Conditional Expectation* (ICE) de Goldstein et al. (2017), sendo que o gráfico abaixo mostra o impacto do estimado na variação do preço no frete estimado para cada uma das observações:

GRÁFICO 15: IMPACTO DO PREÇO NO FRETE ESTIMADO UTILIZANDO ICE NO ALGORITMO DE RANDOM FOREST



Fonte: Elaborado pelo autor

Cada linha se refere à uma observação do conjunto de treinamento e é estimado o valor do frete para cada observação para os diferentes preços possíveis. A linha amarela mostra a curva centralizada em relação ao menor valor do preço, sendo que esta linha mostra a diferença entre o que é estimado e o menor valor possível de preço, possibilitando compreender se a relação entre o frete estimado e o preço é positiva, neutra ou negativa[8].

O gráfico mostra que o algoritmo encontrou uma relação positiva entre preço e frete, indicado pela linha amarela. Conforme o preço aumenta, o valor de frete estimado tende a ser maior. Este resultado vai de acordo com a Hipótese 1, baseado em de Thaler (1980), Kahneman e Tversky (1981) e Azar (2007), na qual a presença de *Relative Thinking* é evidenciada na relação entre frete e preço de produto mesmo após retirar os possíveis efeitos do frete no preço (através do seguro e do peso do produto), sendo que isto se deve ao fato de o consumidor olhar para o

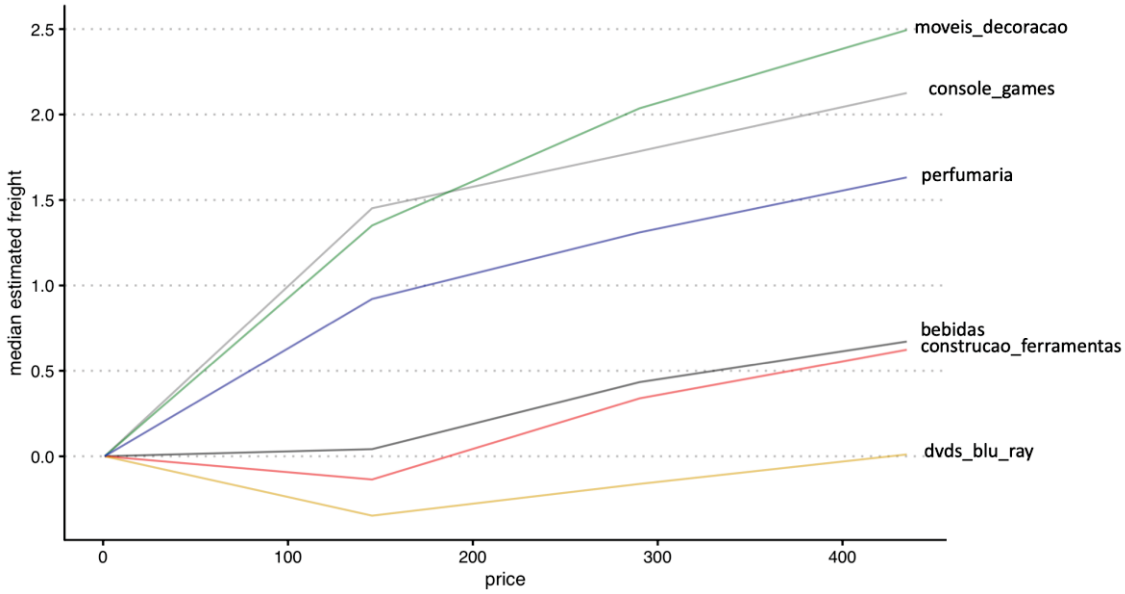
[8] Um ponto importante é desconsiderar regiões com poucas observações em relação ao preço (indicado pelos tracejados no eixo X do gráfico), dado que o algoritmo possui poucas observações para estimar o frete com precisão;

valor do frete de maneira relativa ao valor do preço, fazendo com que fretes altos sejam menos atrativos quando o preço do produto menor e mais atrativos quando o preço do produto é maior.

Outro ponto relevante é que o impacto do preço no frete varia ao longo entre as observações, em alguns casos o impacto é positivo, outros é negativo (indicando que conforme maior o preço, menos a pessoa tende a pagar em frete) e em outros é neutro. Desta maneira foi calculado a mediana do valor de frete estimado pelo ICE para as 6 categorias de produtos[9] para verificar as diferenças de impacto de preço no frete para cada uma destas categorias, com o intuito de testar a Hipótese 2, na qual o impacto do preço sobre o frete é diferente de acordo com a categoria do produto. O gráfico abaixo mostra a relação do frete e da mediana do frete estimado pelo ICE para 6 categorias de produtos distintas:

[9] Foram selecionadas as 3 categorias de produto com a maior mediana do valor de frete estimado pelo ICE e maior número de observações (moveis_decoracao, console_games e perfumaria), além de 3 categorias com a menor mediana do valor de frete estimado pelo ICE e maior número de observações (bebidas, construcao_ferramentas e dvds_blu_ray), com intuito de mostrar a diferença de impacto de preço no frete entre estas categorias de produto;

GRÁFICO 16: MEDIANA DO IMPACTO DO PREÇO NO FRETE ESTIMADO A PARTIR ICE UTILIZANDO O ALGORITMO DE RANDOM FOREST POR CATEGORIA DE PRODUTO



Fonte: Elaborado pelo autor

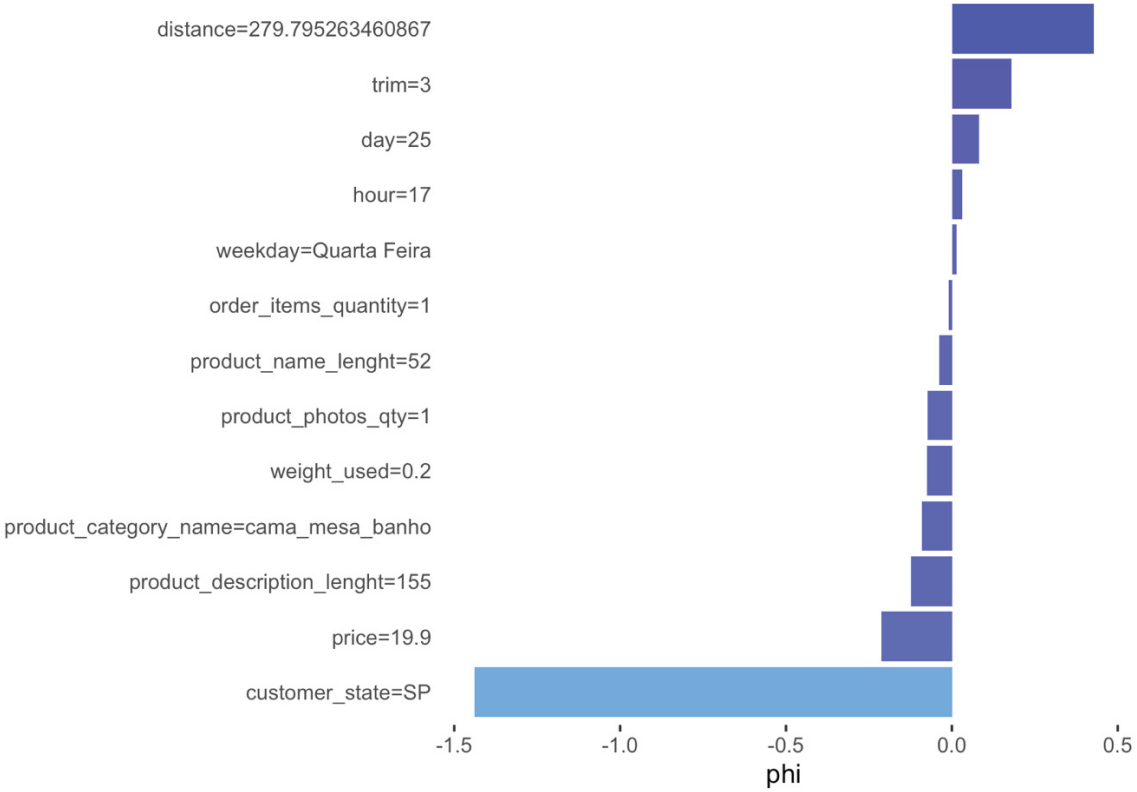
O gráfico mostra que o impacto do preço no frete muda conforma a categoria do produto, em alguns produtos o impacto tende a ser positivo (moveis_decoracao, console_games e perfumaria) e em outros casos tende a ser negativo ou neutro[10] (bebidas, construção_ferramentas e dvd_blu_ray). Este resultado aponta justamente para a Hipótese 2, na qual o impacto do preço no frete difere de acordo com a categoria do produto, dado que o fenômeno de Relative Thinking afeta diferentemente os consumidores a partir de outras categorias de produto, conforme encontrado por por Hirshman et al. (2018).

Além de entender o impacto local do preço no frete, é fundamental compreender o impacto do preço no frete considerando também o impacto das demais variáveis. Para fazer isto foi utilizado *Shapley Values*, sendo que quando aplicamos para uma única observação podemos compreender o impacto calculado

[10] Apesar de, depois de certo preço (150 reais), a mediana começar a subir, observa-se que a mediana do impacto ainda é extremamente baixa (de 2 à 5 vezes mais baixas que as demais categorias)

pelo modelo (*phi*) de cada variável na predição realizada, como pode ser observado no gráfico abaixo:

GRÁFICO 17: SHAPLEY VALUE DE CADA VARIÁVEL PARA UMA OBSERVAÇÃO UTILIZANDO RANDOM FOREST

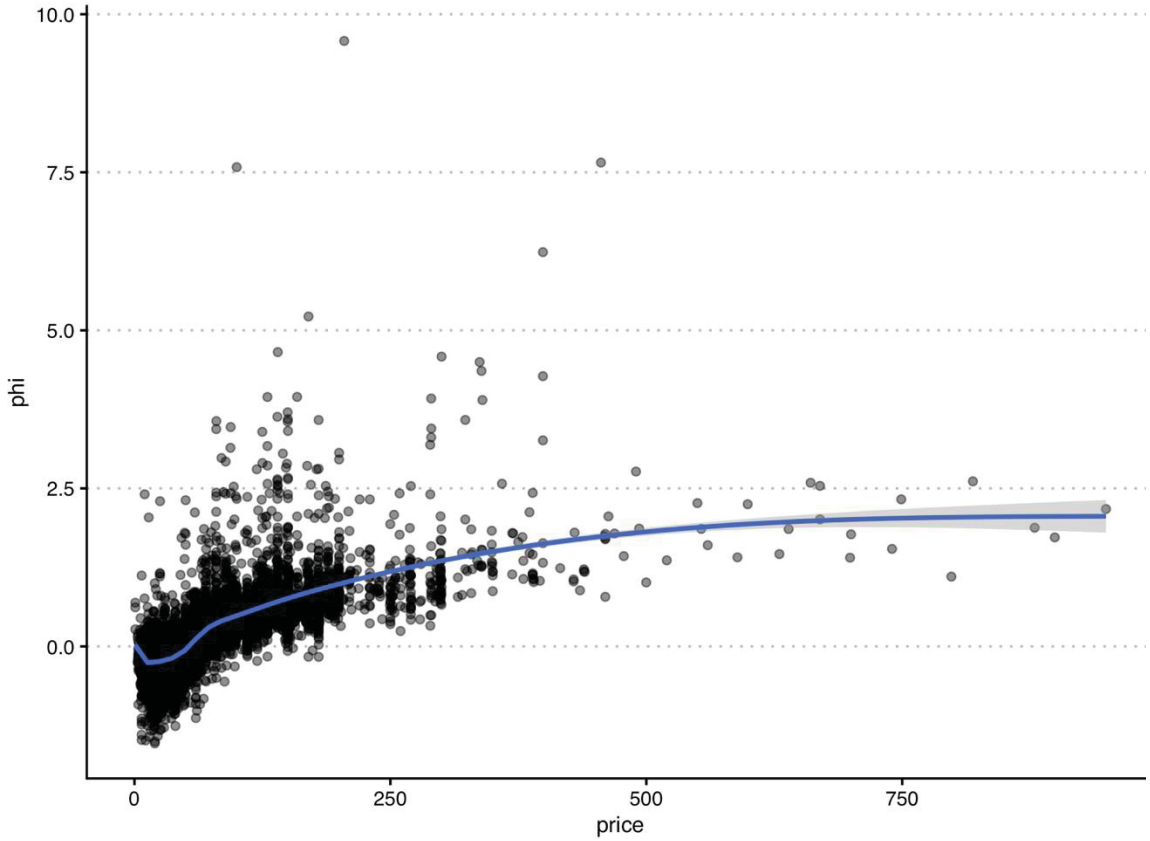


O valor de phi mostra justamente o impacto do valor da variável na predição realizada pelo modelo.

Fonte: Elaborado pelo autor

Para este caso presente na base de dados, observa-se que a distância é de 279.795 quilômetros e que, de acordo com o modelo de *Random Forest*, isto aumenta o valor do frete pago pelo consumidor, contudo o fato de o preço ser R$19,90 está diminuindo o valor do frete pago estimado, além disso, o Estado do consumidor ser São Paulo possui uma impacto negativo extremamente alto no valor de frete pago predito pelo algoritmo. Desta maneira, podemos saber para cada observação qual foi o impacto do preço no valor de frete estimado pelo algoritmo, sendo que o gráfico abaixo mostra justamente a relação de preço pago e o *Shapley Value* relativo ao impacto do preço na predição do frete:

GRÁFICO 18: IMPACTO DO PREÇO NO FRETE UTILIZANDO SHAPLEY VALUES



Fonte: Elaborado pelo autor

O gráfico mostra que quando os preços são mais baixos, o *Shapley Value* tende a ser negativo, indicando ele tende a diminuir o valor do frete estimado, contudo, o *Shapley Value* tende ser positivo e maior, indicando que o valor do frete estimado aumenta. Este resultado mostra que o algoritmo encontrou uma associação positiva entre frete pago e preço, o que vai de acordo com o que foi encontrado com ICE, corroborando com a Hipótese 1.

2.5. CONCLUSÃO

Ao longo do presente trabalho foram testadas duas hipóteses. A primeira delas é referente ao impacto positivo no valor do frete mesmo depois de retirar suas influências através do seguro e impacto do produto, sendo que isto se deveria ao fato de o consumidor olhar para o valor do frete de maneira relativa ao valor do preço, evidenciando o fenômeno de *Relative Thinking* já encontrado por Thaler

(1980), Kahneman e Tversky (1981) e Azar (2007). A segunda delas se refere ao impacto deste viés (*Relative Thinking*) afetar de maneiras distinta os consumidores para categorias de produtos diferentes, sendo que isto foi proposto por Hirshman et al. (2018) a partir dos resultados encontrados por Hasting e Shapiro (2013).

Foram aplicados três modelos para testar estas duas hipóteses: Regressão Linear Múltipla, Árvore de Decisão e *Random Forest.* A escolha destes modelos se deve ao fato de o primeiro ser mais interpretável e paramétrico, o segundo possuir certo grau de interpretabilidade e ser não paramétrico, por fim, o terceiro se deve ao fato de ser flexível para encontrar relações entre variáveis, sendo necessário aplicar neste último técnicas de interpretação de algoritmos como Importância das Variáveis na predição (Fisher, Rudin e Dominici, 2018), Interação entre variáveis (Friedman e Popescu, 2008), *Individual Conditional Expectation* (Goldstein et al., 2017) e *Shapley Values* (Štrumbelj e Kononenko, 2014).

A hipótese 1 foi evidenciada pelo resultado da Regressão Linear Múltipla evidenciou um impacto positivo de preço no frete e estatisticamente significativo (p-valor < 0.05) e pela Árvore de Decisão que utilizou uma quebra relacionada a preços, sendo que quando os preços são maiores o frete estimado pela árvore é maior. Além disso, o resultado encontrado pelo *Random Forest* vai de acordo com esta hipótese, sendo que na Importância de variáveis (Fisher, Rudin e Dominici, 2018) o preço se mostrou importante para prever o valor do frete e através de *Individual Conditional Expectation* (Goldstein et al., 2017) e *Shapley Values* (Štrumbelj e Kononenko, 2014) foi encontrado um impacto positivo do preço no frete mesmo removendo os efeitos de preço no frete pago pelo consumidor através de seguro e peso.

A segunda hipótese foi evidenciada na árvore de decisão, na qual suas quebras indicaram que o impacto do preço depende da categoria do produto. Além disso, a partir

do modelo de *Random Forest* foi calculado, a partir da metodologia de Friedman e Popescu (2008), a interação da variável preço com as demais variáveis na predição do frete, sendo que a maior interação foi entre preço e categoria de produto, apontando novamente para a Hipótese 2. Por fim, a partir da metodologia de *Individual Conditional Expectation* (Goldstein et al., 2017), foi observado que o impacto do preço difere entre categorias de produtos, sendo que para algumas categorias o impacto é positivo (“moveis_decoracao”, “console-games” e

“perfurmaria”) e outras é negativo ou neutro (“bebidas”, “construção_ferramentas” e “dvds_blu_ray”).

Portanto, os resultados obtidos vão em direção às hipóteses levantadas, indicando sim a presença do fenômeno *Relative Thinking* e que seu impacto varia de acordo com a categoria do produto. Além disso, o resultado foi obtido através da utilização de base de dados de uma empresa de *E-commerce*, sendo que anteriormente os trabalhos (Thaler 1980; Kahneman e Tversky, 1981; Azar, 2007; Mowen e Mowen, 1986; Ranyard e Abdel-Nabi, 1993; Frisch, 1993; Moon et al., 1999) que discutiam este fenômeno utilizaram questionários. Além disso, foi utilizada a abordagem de *Machine Learning*, que apesar de ser comumente utilizada em CRM (Ngai, Xiu e Chau, 2009), não é aplicada especificamente para estudar vieses cognitivos, sendo este outro ponto relevante deste trabalho.

Este trabalho apresenta algumas limitações. A primeira delas é que, apesar da evidência experimental do fenômeno de Relative Thinking (Thaler 1980; Kahneman e Tversky, 1981; Azar, 2007; Mowen e Mowen, 1986; Ranyard e Abdel-Nabi, 1993; Frisch, 1993; Moon et al., 1999) e dos controles utilizados em relação ao frete (seguro e peso do produto), não foi estabelecida nenhuma relação de causalidade entre o preço e frete, impossibilitando a inferência de que preços maiores tendem a causar valores maiores de fretes pagos pelos consumidores. Além disso, a renda é uma possível variável de confusão na relação entre preço e frete, sendo que a variável renda não foi levada em consideração no modelo.

Como sugestão para futuros trabalhos fica testar outros algoritmos para compreender se apontam para os mesmos resultados encontrados neste trabalho, assim como mapear quais são as categorias de produtos que aumentam ou diminuem o impacto de *Relative Thinking*. Além disso, incorporar a renda nos modelos para avaliar se a relação de preço e frete persiste e utilizar técnicas de inferência causal para analisar se há uma relação causal entre preço e frete pago pelo consumidor.

3. *ENSAIO 2: MENTAL ACCOUNTING* E *RELATIVE THINKING*: EVIDÊNCIAS NO COMPORTAMENTO DO CONSUMIDOR DE *E-COMMERCE* UTILIZANDO *MACHINE LEARNING*

RESUMO:

Este trabalho utiliza algoritmos de Machine Learning para investigar o impacto variáveis como Qualidade de informação, Qualidade da Entrega (Brown e Jayakody ,2008; Lin, 2007; DeLone e McLean, 2003) e preço do produto (Cao e Gruca, 2004) na satisfação do consumidor. Além disso, é analisado se o modelo de Expectativa e Desconfirmação (Oliver, 1993) afeta a satisfação do consumidor e se o seu impacto é assimétrico (Youjae Yi e Suna La, 2003), evidenciando a presença de Aversão a Perda da Teoria do Prospecto (Kahneman e Tversky, 1979; 1982).

Palavras-chave: *Machine Learning*, Teoria do Prospecto, Desconfirmação

ABSTRACT:

This work uses Machine Learning algorithms to investigate the impact of variables such as Information Quality, Delivery Quality (Brown and Jayakody, 2008; Lin, 2007; DeLone and McLean, 2003) and product price (Cao and Gruca, 2004) in consumer satisfaction. In addition, it is analyzed whether the Expectation and Disconfirmation model (Oliver, 1993) affects consumer satisfaction and whether its impact is asymmetric (Youjae Yi and Suna La, 2003), showing the presence of Loss Aversion of Prospect Theory (Kahneman and Tversky, 1979; 1982).

Palavras-chave: *Machine Learning*, *Prospect Theory, Disconfirmation*

3.1. INTRODUÇÃO

O problema de pesquisa sobre o qual este trabalho se debruça é avaliar quais são os fatores que afetam a satisfação do consumidor em compras online (*e-commerce*). Diversas pesquisas já se debruçaram sobre este problema e encontraram que variáveis como Qualidade do Serviço, Qualidade de informação e Qualidade de Sistema afetam fortemente a satisfação do consumidor (Brown e Jayakody ,2008; Lin, 2007; DeLone e McLean, 2003).

Outra variável importante para a satisfação do consumidor vem do Modelo de Expectativa – Desconfirmação (Oliver, 1993), onde a variável desconfirmação desempenha um papel importante na satisfação do consumidor (Mckinney et al., 2002; Youjae Yi e Suna La, 2003; Szymanski e Henard, 2001; Ryu e Han, 2010). Por fim, Youjae Yi e Suna La (2003) encontrou que a Desconfirmação Negativa possui um impacto maior do que a Desconfirmação Positiva na satisfação do consumidor, sendo isto um indício da Teoria do Prospecto (Kahneman e Tversky, 1979; Kahneman e Tversky, 1992) relacionado a satisfação do consumidor.

A abordagem utilizada em grande parte destes trabalhos se baseia em técnicas de correlação (Ho et al., 2017). Além disso, a grande maioria destes trabalhos é feita utilizando questionários aplicado ao consumidor pelos próprios autores (Brown e Jayakody ,2008; Lin, 2007; DeLone e McLean, 2003; Mckinney et al., 2002; Youjae Yi e Suna La, 2003; Szymanski e Henard, 2001).

O presente trabalho se mostra importante pois ele se diferencia nos pontos citados anteriormente. Primeiramente, ao invés de se utilizar dados de questionário foi utilizado dados de um *e-commerce* brasileiro chamado Olist, sendo ela uma grande rede de *Marketplace* que conecta lojistas a grandes varejistas. Além disso, a metodologia presente neste trabalho não se baseia apenas em técnicas de correção. Dado que a satisfação do consumidor depende de variáveis lineares (Oliver, 2014; Mittal, Ross e Baldasare, 1998) e não lineares (Mittal e Kamakura, 2001), desta maneira utilizou-se algoritmos de *Machine Learning* para analisar a satisfação do consumidor que lidam tanto com relações lineares quanto não lineares.

Os algoritmos utilizados foram *Regressão Logística, Random Forest* e *Artificial Neural Networks*. O primeiro algoritmo foi selecionado devido a sua alta interpretabilidade (James et al., 2013) e os dois últimos devido a sua eficiência para lidar com linearidade e não linearidade presente nas variáveis que afetam a

satisfação do consumidor (Larasati et al., 2012; Leong, 2015), por serem amplamente utilizados pela literatura existente de Gestão de Relacionamento com Cliente (Ngai, Xiu e Chau ,2009) e por serem eficientes para lidar com comportamento do consumidor (Garver, 2002).

Além disso, é analisado a relação da Desconfirmação com a satisfação do consumidor. Para analisar se há impacto assimétrico da Desconfirmação na satisfação do consumidor e se a relação entre as variáveis é representada pela Função Valor da Teoria do Prospecto (Kahneman e Tversky, 1992; Kahneman e Tversky, 1992) foi utilizado *Non-Linear Least Square* (NLLS) para estimar os parâmetros da Função Valor com intuito de identificar se a relação entre as duas variáveis guarda as propriedades de concavidade para ganhos, convexidade para perdas e aversão a perda. Além disso, foi analisado se há diferenças entre os parâmetros de acordo com estado brasileiro, forma de pagamento e tipo de produto comprado. Por fim, utilizando *Non-Linear Least Square* (NLLS), foi analisado se os três algoritmos encontraram uma relação assimétrica e que respeita a forma funcional da Função Valor proposta por Kahneman e Tversky (1992).

A seção 2 traz a revisão bibliográfica, mostrando quais são os fatores que afetam a satisfação do consumidor, o modelo de Expectativa-Desconfirmação, o uso de Machine Learning para entender o comportamento do consumidor, acerca da Teoria do Prospecto e as hipóteses analisadas no presente trabalho. A seção 3 apresenta as metodologias utilizadas. A seção 4 mostra os resultados obtidos. Por fim, a seção 5 traz as conclusões e considerações finais do trabalho.

3.2. REVISÃO BIBLIOGRÁFICA

3.2.1. FATORES QUE AFETAM A SATISFAÇÃO DO CONSUMIDOR EM E-COMMERCE

Existem diversas definições possíveis de satisfação, contudo, será utilizado a definição fundamentada por Spreng et al. (1996), em que a satisfação geral é definida como um estado de afetividade que representa a reação emocional do consumidor ao longo de sua experiência de compra (seja ela online ou não). A satisfação do consumidor é uma consequência da experiência que este consumidor teve ao comprar determinado produto, contudo, quando se trata de consumo na

internet (e-commerce) é necessário considerar que a quantidade de informações é limitada, sendo que justamente devido a este fator é preciso analisar outras variáveis para compreender a satisfação do consumidor em compras online.

A partir do modelo clássico de DeLone e McLean (2003), as principais variáveis consideradas relevantes para a satisfação do consumidor são Qualidade da Informação, Qualidade do Sistema e Qualidade do Serviço. Qualidade da Informação pode ser definido como as características da informação, isto é, se as mesmas são úteis, completas, atualizadas etc (DeLone e McLean, 2003). Ainda na definição de DeLone e McLean (2003), pode-se definir a Qualidade do Sistema como um reflexo da usabilidade, disponibilidade, adaptabilidade, confiabilidade e tempo de resposta do sistema. Por fim, os autores definem a Qualidade do Serviço como o suporte geral dado pelo provedor de serviço, sendo que Zeithmanl et al. (2002) definem, mais especificamente, como as facilidades oferecidas para facilitar compras e entrega do produto.

Brown e Jayakody (2008), utilizando múltiplas regressões lineares com dados de um questionário respondido por alunos de pós graduação, analisaram a relação entre a satisfação do consumidor e sete medidas: Qualidade do Sistema, Qualidade da Informação, Qualidade de Serviço, Confiança, Incentivos de Lealdade, Utilidade Percebida, Satisfação do Consumidor e Intenção de Continuar Comprando. Os autores encontraram impacto significativo de Qualidade de Serviço e Utilidade Percebida, além disso, observaram que a Utilidade Percebida é influenciada pela Qualidade da Informação e a Confiança, sendo essa última significativamente influenciada pela Qualidade do Sistema e pela Qualidade do Serviço.

Lin (2007), utilizando modelagem de equações estruturais, encontrou influência na satisfação do consumidor da Qualidade do Sistema (Através do design do website e a interatividade), da Qualidade da Informação (através da informatividade e segurança) e da Qualidade do Serviço (através da capacidade de resposta e confiança), o que reforça os resultados dos outros trabalhos já mencionados.

Outras variáveis desempenham um papel importante na determinação da satisfação do consumidor. Cao e Gruca (2004) avaliaram as classificações dos consumidores antes e depois das compras para três lojas virtuais grandes de livros e encontraram que as maiores classificações estavam ligadas a preços maiores, tanto

nas classificações antes das compras quanto nas classificações depois das compras.

Por fim, Lin et al. (2011) aplicaram questionários para alunos de graduação para medir o impacto de diversas variáveis na satisfação do consumidor (Qualidade da Informação, Qualidade do Sistema, Qualidade do Serviço, Qualidade do Produto, Qualidade da Entrega e Preço Percebido). Utilizando regressões múltiplas, encontraram que a variável mais impactante na satisfação do consumidor foi a Qualidade da Entrega, seguido pela Qualidade do Produto. Este resultado está de acordo com o trabalho de Ahn et al. (2004) que utilizou Modelagem Estrutural a partir de dados de uma pesquisa online baseado em seis empresas online da Korea. Eles encontraram que as variáveis mais importantes para explicar a utilidade percebida pelo consumidor é primeiramente Qualidade da Informação, sendo seguida por Qualidade do Produto e Qualidade da Entrega. Estes resultados mostram que os elementos cruciais que as empresas de e-commerce devem se concentrar é na provisão de informação ao consumidor e na execução da entrega após a compra, sendo que estes são justamente os serviços que o e-commerce deve oferecer para incentivar o consumidor a comprar (Cao e Gruca, 2004).

3.2.2. MODELO DE EXPECTATIVA E DESCONFIRMAÇÃO

Segundo Oliver (1993), o Modelo de Expectativa e Desconfirmação trata da dinâmica de três elementos: a expectativa, a performance real e a desconfirmação. Esta última se refere à percepção que surge quando um indivíduo possui expectativas acerca de um produto ou serviço e estas expectativas são confrontadas com a performance real do mesmo, podendo gerar uma desconfirmação positiva (caso a performance real seja maior do que a expectativa) ou desconfirmação negativa (caso a performance real seja menor do que a expectativa).

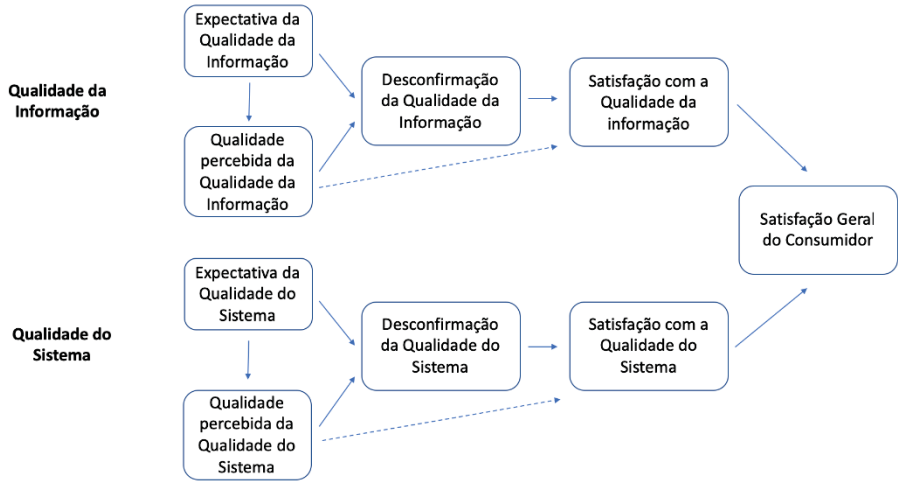
Já foram detectadas influências de desconfirmação em diversos setores, como por exemplo em agências de viagem online (McKinney et al. 2002), provedores de serviços de aplicativos (Susarla et al. 2006), internet móvel (Thong et al., 2006), uso de tecnologias da informação (Bhattacherjee, 2001), jogos online (Park e Lee, 2011) e serviços de e-commerce (Ho et al. 2017). Segundo Ho et al. (2017), grande parte dos estudos para analisar o efeito da desconfirmação no comportamento do consumidor se restringe a técnicas de correlação. Uma

metodologia baseada em técnicas de correlação extremamente utilizada para análise da satisfação do consumidor é *importance – performance analysis*, mas Matzler et al. (2004) sustentam que as implicações gerenciais derivadas desta abordagem em relação à satisfação do consumidor são enganosas. Um dos primeiros modelos a introduzir uma abordagem diferente foi proposto por Ho et al. (2017). Os autores utilizaram dados disponibilizados por um site de e-commerce para avaliar o impacto da desconfirmação na decisão de compartilhar ou não a satisfação. Eles trabalharam com uma abordagem hierárquica Bayesiana, além de desenvolverem uma seção de habilidade de predição do modelo (através da acurácia), o que se diferiu demasiadamente do restante da literatura.

Mckinney et al. (2002) propõem um modelo de Expectativa-desconfirmação focado na Qualidade da Informação e na Qualidade do Sistema, em que os consumidores fazem inferência acerca da qualidade do produto baseado (1) na informação oferecida pelo vendedor e (2) nos elementos de design do site que promovem melhor navegação. Para os autores, a satisfação do consumidor depende de três elementos: a expectativa, a desconfirmação e a performance percebida. A *expectativa*, definida pelos autores a partir da definição de Teas (1993) e Szajna e Scamell`s (1993), é aquilo que o consumidor espera que o produto fará, isto é, são as crenças do consumidor acerca dos atributos e/ou performance do produto (Sprend et al., 1996). Contrastando com a expectativa do consumidor, existe a performance efetiva do produto que será avaliada pelo consumidor, sendo esta denominada de *performance percebida*. Por fim, há a desconfirmação, que é definido pelos autores como os julgamentos subjetivos do consumidor ao comparar as expectativas com a performance real.

Assim, para Mckinney et al. (2002), a satisfação do consumidor se deve justamente a essa dinâmica entre expectativa e performance percebida, a qual pode gerar a desconfirmação, que afeta diretamente a satisfação do consumidor: quanto maior a desconfirmação negativa, isto é, uma performance pior do que a expectativa, menor será a satisfação do consumidor. Relacionando essa ideia para os dois níveis de qualidade (Informação e Sistema), os autores propõem o diagrama abaixo:

QUADRO 1 – MODELO DE EXPECTATIVA - DESCONFIRMAÇÃO



Fonte: Adaptado de McKinney et al. (2002)

Para confirmar este modelo empiricamente os autores utilizaram Modelagem de Equações Estruturais a partir de dados obtidos através de questionário. A partir dos resultados, os autores observaram que as expectativas e a desconfirmação são medidas importantes ao analisar a satisfação do consumidor, além de mostrarem que a Qualidade da informação e a Qualidade do Sistema desempenham um papel essencial na satisfação do consumidor.

Além disso, Youjae Yi e Suna La (2003) encontraram que o impacto da desconfirmação sobre a satisfação do consumidor pode ser assimétrica. Utilizando Modelagem de Equações estruturais, os autores mostraram que a formação da satisfação do consumidor é moderada pela confiança do mesmo em suas expectativas: o efeito da performance na satisfação do consumidor é maior em grupos mais confiantes do em grupos menos confiantes. Por outro lado, os autores mostram que, semelhante ao resultado de grande parte da literatura da área, as expectativas não possuem um efeito direto significativo na satisfação dos consumidores. Além disso, os autores encontraram que Desconfirmação negativa possui um impacto muito maior do que Desconfirmação positiva na satisfação dos consumidores, o que está de acordo com a Teoria do Prospecto proposta por Kahneman e Tversky (1979).

Outros autores encontraram impacto significativo da desconfirmação na satisfação do consumidor. Szymanski e Henard (2001), através de uma meta-análise de evidências empíricas, detectaram uma forte relação entre desconfirmação e

satisfação do consumidor. Ryu e Han (2010) utilizaram Modelagem de Equações Estruturais para analisar serviços de restaurantes e encontraram que fatores como a aparência, iluminação, fatores sociais e outras variáveis afetam severamente a desconfirmação e que essa, por sua vez, possui uma influência na satisfação do consumidor.

3.2.3. MACHINE LEARNING E SATISFAÇÃO DO CONSUMIDOR

A literatura mostra que ao analisar a performance de atributos na satisfação geral do consumidor existem relações lineares destes atributos com a satisfação (Oliver, 2014; Mittal, Ross e Baldasare, 1998) e relações que são não lineares (Mittal e Kamakura, 2001). Desta maneira, métodos estatísticos tradicionais tendem a não ter boa performance na predição de satisfação de consumidor, visto que precisam assumir a hipótese de linearidade (Larasati et al., 2012). Uma alternativa para esta metodologia são algoritmos de *Machine Learning*, como por exemplo, *Artificial Neural Networks* (ANN) e *Decision Tree.*

O uso de *Artificial Neural Networks* (ANN) é relevante para lidar com a satisfação do consumidor pelo fato deste algoritmo conseguir encontrar fatores não compensatórios, sendo capaz de detectar relações não lineares e lineares das diversas variáveis com a satisfação do consumidor (Leong, 2015). Além disso, o algoritmo consegue lidar com multicolinearidade entre as variáveis e analisar qualquer tipo de distribuição (Lin, 2007b). Ainda, a ANN não necessita dos pressupostos exigidos nas metodologias tradicionais, como por exemplo linearidade, normalidade das variáveis ou homocedasticidade (Lee, Leong, Hew e Ooi, 2013). Algumas pesquisas comparam o uso de ANN em relação a métodos estatísticos, e mostram que a primeira metodologia possui melhores resultados do que a segunda (West, Brockett e Golden, 1997; Gronholdt e Martensen, 2005). Contudo, há uma grande limitação nesta metodologia: o resultado gerado pelo algoritmo é incompreensível, isto é, o algoritmo é uma “caixa preta” (Bounsaythip e Runsala, 2001).

Além da *Artificial Neural Networks*, outras técnicas de mineração de dados se mostram eficientes por não precisarem dos pressupostos necessários nestas

metodologias tradicionais. Aém de ANN, tem sido proposto utilizar o algoritmo de *Decision Tree* para analisar o comportamento do consumidor (Garver, 2002). Uma grande vantagem do algoritmo de *Decision Tree* é que o resultado é fácil de compreender e não é afetado por dados faltantes (Bounsaythip e Runsala, 2001).

Analisando especificamente a área de Retenção de Consumidor (onde a satisfação do consumidor é analisada para as áreas de Marketing *one-to-one* e Programas de Lealdade), o algoritmo de *Decision Tree* e *Neural Networks* é aplicado conjuntamente em diversos trabalhos como, por exemplo, Hung et al. (2006), Koh e Chan (2002), Mozer et al. (2000) e Smith et al. (2000). Além disso, *Neural Networks* foi aplicado conjuntamente com *Random Forests* (uma abordagem baseada em árvores de decisão) no trabalho de Buckinx e Poel (2005) e Larivie'Re e Poel (2005).

Segundo (Lasarati et al., 2012), os trabalhos que aplicam redes neurais em pesquisas relacionadas ao consumidor são divididos em duas grandes áreas cujos objetivos são diferentes: determinar qual o fator mais crítico que influencia a satisfação do consumidor e predizer a satisfação geral do consumidor. Tsaur et al. (2002) utilizaram ANN e regressão logística para medir a importância dos aspectos de serviço na satisfação do consumidor para nove hotéis internacionais. Através da comparação dos resultados, os autores concluiram que ANN se mostra melhor para lidar com não linearidades e com maior poder de previsão. Deng et al. (2008) utilizou ANN para analisar o impacto da qualidade do serviço em um hotel de Taiwan através de um questionário aplicado aos consumidores, em que o objetivo foi analisar quais são as variáveis mais importantes para predizer a satisfação do consumidor.

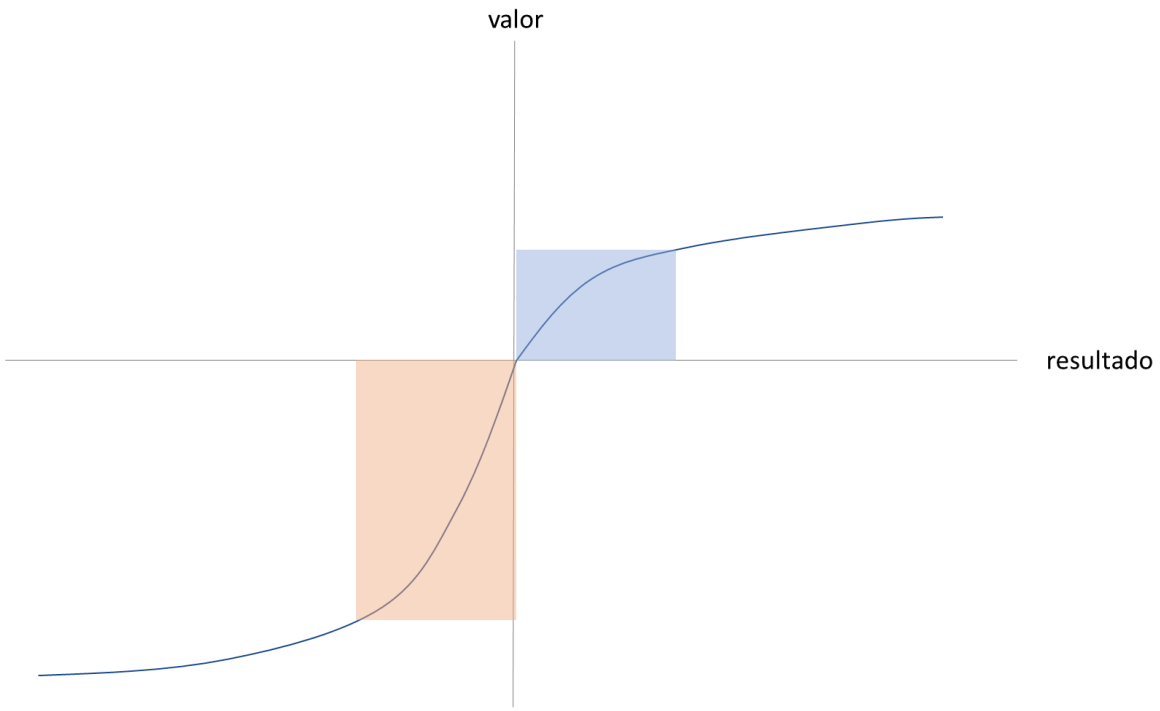
Behara et al. (2002) utilizaram redes neurais a partir de um questionário aplicado aos consumidores com o objetivo de encontrar a melhor configuração da ANN para predizer a satisfação do consumidor e avaliar o impacto da qualidade do serviço Além disso, os autores encontraram que a modelagem que envolvia a diferença entre expectativa e percepção do consumidor possuiu uma performance superior do que apenas percepção ou expectativa, sendo que tal resultado vai de acordo com o Modelo de Expectativa – Desconfirmação discutido na seção anterior.

3.2.4. TEORIA DO PROSPECTO

Kahneman e Tversky (1979), no trabalho intitulado de "*Prospect Theory: An Analysis of Decision under Risk",* propuseram uma alternativa à Teoria da Utilidade Esperada para explicar a maneira pela qual as pessoas tomam decisões em situações de risco. Na teoria intitulada de "Teoria do Prospecto", os autores propõem que a tomada de decisão passa por duas etapas: a primeira é a "edição" e a segunda é a "avaliação". A fase de "edição" consiste em uma análise preliminar na qual as pessoas tendem a simplificar as opções que são oferecidas a elas. Uma das operações que ocorre nessa fase é a "codificação", em que as pessoas tendem a avaliar os resultados das opções oferecidas como valores relativos a um ponto de referência neutro e não como um resultado absoluto. Na segunda etapa ("avaliação"), as opções são avaliadas após terem sido "editadas", sendo que esta avaliação é explicada pela *Value Function* e *Weighting Function,* sendo a primeira o foco do presente trabalho.

No que tange a *Value Function,* existem três aspectos principais. O primeiro deles é que os valores financeiros das opções oferecidas (isto é, os resultados) são definidos como desvios a partir de um ponto de referência considerado neutro. O segundo é que esta função é côncava para ganhos (em relação ao ponto de referência) e convexa para perdas. Por fim, o terceiro ponto se refere à esta curva ser mais íngreme para perda do que para ganhos, isto é a derivada desta função para um ganho de *x* é menor do que a derivada desta função para uma perda de *x.* Baseado nestes aspectos, no gráfico abaixo é mostrada a curva da *Value Function* que satisfaz estas três propriedades, em que para resultados positivos a função é côncava e para resultados negativos a função é convexa. Assim, para uma dada variação *z* a partir do ponto de referência observa-se que o impacto no valor (em módulo) é maior quando a variação é negativa, indicando que a função é mais íngreme para perdas do que para ganhos:

GRÁFICO 19: CURVA DA FUNÇÃO VALOR PROPOSTA PELA TEORIA DO PROSPECTO



Fonte: Elaborado pelo autor com base em Kahneman e Tversky (1979)

Kahneman e Tversky (1992) desenvolvem a Teoria do Prospecto Cumulativa, trazendo alterações principalmente para *Weighting Function*. Na Teoria do Prospecto (Kahneman e Tversky, 1979) os pesos eram atribuídos através de uma transformação monotônica na probabilidade de cada uma das opções separadamente, já na Teoria do Prospecto Cumulativa a transformação é aplicada em toda função de distribuição cumulativa das probabilidades. De acordo com Kahneman e Tversky (1992), esta alteração possibilita a aplicação da Teoria do Prospecto para qualquer número finito de opções e é possível estender para variáveis contínuas.

Em relação à *Value Function*, Kahneman e Tversky (1992) assumem que (i) a função é côncava para ganhos e convexa para perdas; (ii) a função é mais íngreme para perdas do que para ganhos, semelhante ao que foi proposto por Kahneman e Tversky (1979). De acordo com autores, o primeiro pressuposto pode ser explicado pela sensibilidade decrescente, isto é, o impacto das mudanças na *Value Function* é menor conforme mais distante do ponto de referência. O segundo pressuposto é

explicado pelo princípio de aversão à perda, na qual as pessoas têm a impressão de que perdas são maiores do que ganhos correspondentes.

Levando esses pressupostos em consideração, Kahneman e Tversky (1992) propuseram a seguinte forma funcional para a *Value Function*:

$$v(x) = \begin{cases} x^{\alpha}, & se\ x \geq 0 \\ -\lambda(-x)^{\beta}, & se\ x < 0 \end{cases} \qquad (7)$$

O que se observa é que para valores de *x* maiores que zero (isto é, resultados positivos a partir do ponto de referência), a função é dada por x^{α}, onde α indica a curvatura da *Value Function* para ganhos, sendo que 0 < α < 1 para que a função seja côncava para ganhos e indicando a sensibilidade decrescente para ganhos. Para valores de *x* menores que zero (isto é, resultados negativos a partir do ponto de referência), a função é dada por $-\lambda(-x)^{\beta}$, onde β indica a curvatura da *Value Function* para perdas, sendo que para que a função seja convexa conforme teorizado por Kahneman e Tversky (1979; 1992) é necessário que 0 < β < 1, fazendo com que a função apresente sensibilidade decrescente para perdas. Além disso, o termo λ mede o grau de aversão a perdas, mostrando justamente quanto a *Value Function* é íngreme para perdas do que para ganhos.

Barberis (2013) reúne uma vasta literatura relacionada à aplicação da Teoria do Prospecto nos últimos 30 anos e mostra que existem grandes desafios na aplicação desta teoria. O autor defende que isso se deve ao fato de que as pessoas percebem ganhos e perdas a partir de um ponto de referência, ideia central da Teoria do Prospecto. A grande dificuldade é definir qual o ponto de referência a ser utilizado, fazendo com que muitas aplicações desta teoria sejam impraticáveis. Justamente por isso são poucos os trabalhos de aplicação da Teoria do Prospecto e a maior parte deles está restrita às áreas financeiras e de seguro (Barberis, 2013)

No presente trabalho, a definição adotada de ponto de referência será baseada em Köszegi e Rabin (2006, 2007, 2009), nos quais os autores propõem que o ponto de referência deve ser escolhido baseado nas expectativas racionais que os consumidores possuem a partir dos resultados obtidos em situações passadas semelhantes.

Booij et al. (2010) reúnem diversas estimações dos parâmetros da *Value Function* até 2010, isto é, os parâmetros α, β e λ. A tabela abaixo mostra os valores

dos parâmetros estimados para a *Value Function* reunidos por Booij et al. (2010), assim como valores de parâmetros estimados pelos próprios autores:

TABELA 3: PARÂMETROS DA FUNÇÃO VALOR ESTIMADOS PELA LITERATURA

Autores	**α**	**β**	**λ**
Tversky e Kahneman (1992)	0.88	0.88	2.25
Camerer e Ho (1994)	0.22		
Wu e Gonzalez (1996)	0.50		
Fennema e van Assen (1998)	0.39	0.84	
Gonzalez e Wu (1999)	0.49		
Abdellaoui (2000)	0.89	0.92	
Donkers et al. (2001)	0.61	0.61	
Schmidt e Traub (2002)			1.43
Etchart-Vincent (2004)		0.97	
Abdellaoui et al. (2005)	0.91	0.96	
Tu (2005)	0.68	0.74	3.2
Fehr-Duda et al. (2006)	1.01	1.05	
Abdellaoui et al. (2007)	0.72	0.73	2.54
Andersen et al. (2006)	0.81	0.80	1.07
Harrison and Rutström (2009)	0.71	0.80	1.07
Abdellaoui et al. (2008)	0.86	1.06	2.61
Booij et al. (2010)	0.86	0.83	1.58

Fonte: Elaborado pelo autor com base em Booji et al. (2010)

Segundo Booji et al. (2010), é importante observar nestes resultados que, em geral, os parâmetros α e β são menores do que 1, indicando justamente a sensibilidade decrescente proposta por Kahneman e Tversky (1979; 1982). Também, na maioria dos estudos o parâmetro β é maior que α, indicando que perdas são avaliadas mais linearmente do que ganhos, fazendo com que as pessoas tendam a se tornar menos sensíveis a ganhos mais rapidamente do que a perdas. Por fim, há uma variação alta nas estimativas do parâmetro que mede aversão à perda (λ), contudo ele sempre é maior que 1.

Youjae Yi e Suna La (2003) mostram que a Teoria da Expectativa e Desconfirmação possui diversas semelhanças com a Teoria do Prospecto.

Primeiramente ambas dependem de um ponto de referência, definido pelas expectativas das pessoas. Em segundo lugar, a percepção das pessoas de satisfação é dada a partir de ganhos e perdas a partir deste ponto de referência, segundo a Teoria do Prospecto. De forma semelhante, na Teoria da Expectativa e Desconfirmação os ganhos a partir do ponto de referência são descritos como Desconfirmação Positiva (quando a performance é maior que a expectativa) e perdas são descritas como Desconfirmação Negativa (quando a performance é menor que a expectativa). Em terceiro lugar, na Teoria do Prospecto, a *Value Function* mede o valor subjetivo ou utilidade e a Teoria da Desconfirmação mede justamente a satisfação do consumidor, que está intimamente relacionada com a Utilidade.

Youjae Yi e Suna La (2003), utilizando modelagem de equações estruturais, encontraram um impacto assimétrico da Desconfirmação na satisfação do consumidor, em que a Desconfirmação negativa possui um impacto maior do que a Desconfirmação positiva, apontando justamente para o fenômeno de aversão à perda. O mesmo resultado foi encontrado por Schifferstein et al. (1999) e Cheung e Lee (2009), apontando para o efeito da aversão a perda no contexto de desconfirmação e satisfação do consumidor.

Com base na literatura apresentada, nota-se que a Qualidade da Informação, Qualidade do Produto, Qualidade da Entrega e preço do produto são relevantes para a satisfação do consumidor. Também foi identificado que o modelo de Expectativa-Desconfirmação exerce impacto na satisfação do consumidor, sendo que em casos que a performance é maior ou igual a expectativa, o consumidor fica satisfeito e em casos onde a performance é inferior a expectativa, o consumidor fica insatisfeito. Por fim, também é apresentado na literatura que o impacto da Desconfirmação é assimétrico, isto é, a Desconfirmação Negativa exerce um impacto maior do que a Desconfirmação Positiva, apontado para o conceito de Aversão à Perda apresentado na Teoria do Prospecto (Kahneman e Tversky, 1979) e a forma funcional da Função Valor (Kahneman e Tversky, 1992). Desta maneira, as hipóteses a serem testadas no presente trabalho são:

Hipótese 1: A Qualidade da Informação afeta a satisfação do consumidor (medido pela quantidade de caracteres no título, quantidade de caracteres na descrição e quantidade de fotos);
Hipótese 2: A Qualidade da Entrega afeta a satisfação do consumidor (medido através da variável desconfirmação, a qual compara a o tempo de entrega estimado com o tempo de entrega real);
Hipótese 3: O preço do produto afeta a satisfação do consumidor;
Hipótese 4: A desconfirmação possui um impacto positivo e relevante na satisfação do consumidor;
Hipótese 5: A relação da Desconfirmação com a satisfação do consumidor segue a forma funcional da Função Valor proposta por Kahneman e Tvserky (1992);
Hipótese 6: Os parâmetros de concavidade, convexidade e aversão à perda variam conforme o Estado do consumidor, método de pagamento e categoria de produto comprado;

Serão utilizados três algoritmos (Regressão Logística, *Random Forest* e *Artificial Neural Networks*) para entender o impacto das variáveis presentes nas hipóteses na satisfação do consumidor. Por fim, para analisar o impacto da variável Desconfirmação e estimar os parâmetros relacionados à Função Valor (Kahneman e Tversky, 1992), foi utilizado *Non Linear Least Squares* (NLLS), para comparar os parâmetros estimados entre estados brasileiros, métodos de pagamento e categoria de produto.

3.3. METODOLOGIA

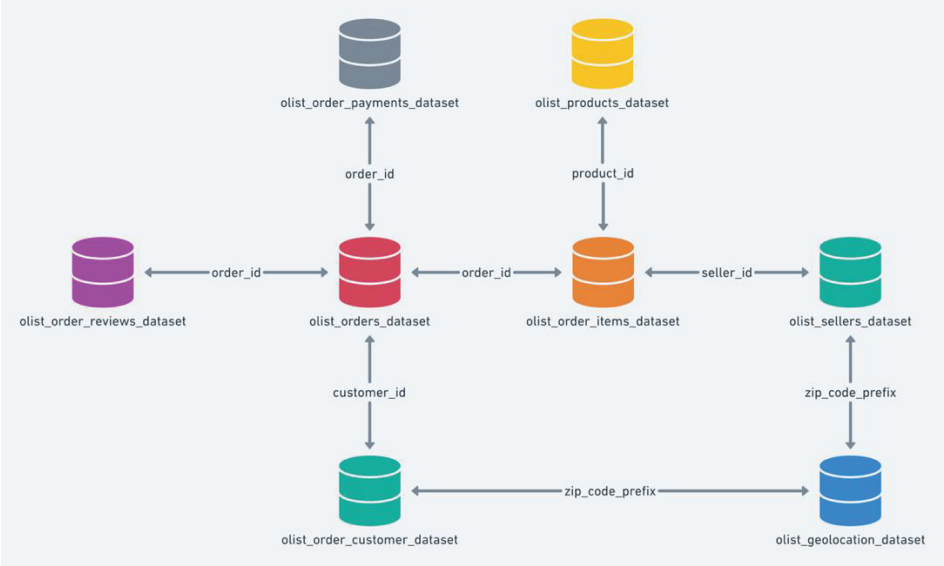
3.3.1. BASE DE DADOS E APLICAÇÃO DE ALGORITMOS

Foram utilizados os dados de uma empresa de *E-commerce* brasileira chamada *Olist,* sendo que as bases de dados estão disponíveis no Kaggle. Primeiramente, a base de dados contendo os itens comprados (*olist_order_items_dataset*) foi sumarizada com o intuito de retirar pedidos sequenciais, isto é, casos em que a pessoa comprou mais unidades do mesmo produto. Dessa forma, ficou na base somente o preço individual do produto (*price*), o

preço do frete (freight_value) e a quantidade de itens comprados (*order_items_quantity*).

Em seguida, as bases de dados foram unidas seguindo a instrução contida no próprio Kaggle, como pode ser observada abaixo:

FIGURA 1: INTEGRAÇÃO DE BASE DE DADOS DA OLIST



Fonte: KAGGLE

A partir da base de dados de pedidos (*olist_orders_dataset*) foram trazidas as informações de *review* dos pedidos (*olist_order_reviews_dataset*), dados dos consumidores (*olist_order_customer_dataset)*, dados de itens comprados (*olist_order_items_dataset*), dados dos produtos (*olist_products_dataset*) e dados de pagamento (*olist_order_payments_dataset*).

Após juntar as bases de dados, as colunas mantidas foram: nota dada pelo consumidor (*review_score*), *status* do pedido (*order_status*), preço do produto (*price),* valor do frete (*freight_value*), quantidade de itens comprados (*order_items_quantity*), tamanho do nome do produto (*product_name_lenght*), tamanho da descrição do produto (*product_description_lenght*), quantidade de fotos (*product_photos_qty*), data prevista de entrega (*order_estimated_delivery_date*), data de entrega (*order_delivered_customer_date*), data de resposta da survey (*review_answer_timestamp*), tipo de pagamento (*payment_type*), quantidade de

parcelas (*payment_installments*), momento da compra (*order_purchased_timestamp)* e utilização de voucher (*used_voucher*).

A partir da data de resposta da *survey* (*review_answer_timestamp*) foi criado um campo referente à hora que a *survey* foi respondida pelo consumidor (*hour_survey_answer*). Além disso, utilizando os campos referentes à data prevista de entrega (*order_estimated_delivery_date*), data de entrega (*order_delivered_customer_date*) e momento da compra (*order_purchased_timestamp*) foi criado o campo relacionado à desconfirmação (*disconfirmation*) através da subtração dos dias demorados para entrega (data de entrega menos o momento de compra) dos dias estimados de entrega (data prevista de entrega menos o momento da compra), fazendo que em casos que a entrega atrasou, o campo *disconfirmation* seja negativo e em casos em que a entrega chegou antes do previsto o campo *disconfirmation* seja positivo.

A variável referente à nota dada pelo consumidor (*review_score*) se refere a uma variável discreta (indo de 1 à 5), desta maneira ela foi transformada em uma variável binária para utilização de algoritmos de classificação binários. Para isto os casos em que o *review_score* são 4 e 5, o campo *review_score* foi denominado "s*atisfeito"* e em casos nos quais o *review_score* são 1 e 2, o campo *review_score* foi denominado "*insatisfeito".* Por fim, os casos em que o *review_score* equivale a 3 foram removidos da base de dados. Esta quebra se justifica pelo baixo número de observações referentes às notas 1 (10603 observações),2 (3257 observações) e 3 (8378), e alto número de observações presentes nas notas 4 (19282) e 5 (57999). Contudo, apesar de diminuir o desbalanceamento entre as *labels* relacionadas a nota dada pelo consumidor, ainda há um desbalanceamento entre as labels de "satisfeito" (77281 observações) e "insatisfeito" (13860 observações).

Após estas transformações na base foram removidos os campos: data prevista de entrega (*order_estimated_delivery_date*), data de entrega (*order_delivered_customer_date*), momento da compra (*order_purchased_timestamp*) e data de resposta da *survey* (*review_answer_timestamp*).

Após o tratamento da base de dados, havia 91.140 observações e 13 variáveis. Das variáveis presentes na base de dados final, 3 variáveis eram categóricas: satisfação do consumidor (*review_score*), tipo de pagamento

(*payment_type*) e utilização de *voucher* (*used_voucher*). As demais variáveis são numéricas, sendo que abaixo pode-se observar a estatística descritiva delas:

TABELA 4: ESTATÍSTICA DESCRITIVA DE VARIÁVEIS NUMÉRICAS PARA AVALIAR A SATISFAÇÃO DO CONSUMIDOR

Numerical Variables	Mean	Stdev	Median	Minimum	Maximum	1.Quartile	3.Quartile
payment_installments	2,974529	2,752984	2	0	24	1	4
price	123,7889	188,0082	78	0,85	6735	40	139
freight_value	20,07056	15,77464	16,32	0	409,68	13,14	21,19
order_items_quantity	1,099891	0,456358	1	1	20	1	1
product_name_lenght	48,90612	9,97874	52	5	76	43	57
product_description_lenght	785,7633	649,5562	602	4	3992	348	985
product_photos_qty	2,238252	1,741986	2	1	20	1	3
hour_survey_answer	13,13569	7,224893	14	0	23	10	19
disconfirmation	11,38907	10,22241	12	-189	146	7	16

Fonte: Elaborado pelo autor

3.3.2. ALGORITMOS DE *MACHINE LEARNING*

Para testar as hipóteses propostas, será utilizada uma abordagem de Aprendizagem Supervisionada, mais especificamente uma abordagem de Classificação, em que a variável dependente é a satisfação do consumidor (*review_score*), a qual é uma variável binária (consumidor satisfeito ou consumidor insatisfeito). Todas as hipóteses levantadas são relacionadas aos impactos das diferentes variáveis na satisfação do consumidor

James et al. (2013) apresentam critérios que diferenciam algoritmos de *Machine Learning*. O primeiro deles é se estes algoritmos são paramétricos ou não-paramétricos. Em geral, algoritmos paramétricos assumem explicitamente uma forma funcional de tal maneira que o modelo calcula apenas os parâmetros. Há o risco, contudo, de que os pressupostos utilizados pela função assumida (por exemplo, linearidade) não ocorram realmente nos dados analisados, fazendo com que a qualidade do modelo seja baixa. Os modelos não-paramétricos não assumem nenhuma forma funcional específica e, em geral, dependem de poucos pressupostos.

A segunda maneira de diferenciar estes algoritmos é baseado no *trade-off* de flexibilidade e acurácia destes modelos. Em geral, modelos com alta flexibilidade

(geralmente, modelos não paramétricos) são difíceis de interpretar e modelos com baixa flexibilidade (geralmente, modelos paramétricos) possuem maior interpretabilidade.

Por fim, a terceira diferenciação que pode ser utilizada é a motivação pela qual o modelo será utilizado: inferência ou predição. No caso da inferência, o objetivo é compreender como Y está sendo afetado pelas variáveis dependentes, desta maneira é importante que o modelo seja interpretável. No caso da predição, o objetivo é apenas prever corretamente Y, desta maneira tendem a ser utilizados algoritmos com alta flexibilidade e baixa interpretabilidade.

Os algoritmos escolhidos no presente trabalho são baseados nestes três critérios. O primeiro deles é a Regressão Logística, o qual é paramétrico, possui alta interpretabilidade e é apropriado para fazer inferência estatística (James et al., 2013). Desta maneira, este primeiro modelo permite entender as relações entre as variáveis de maneira simples através dos parâmetros estimados, fazendo com que a acurácia do modelo seja baixa caso os pressupostos não sejam respeitados pelos dados utilizados.

O segundo algoritmo utilizado foi o *Random Forest*, o qual é não paramétrico e possui baixa interpretabilidade. O resultado deste algoritmo geralmente possui uma boa acurácia de predição, contudo seu resultado tende a ser difícil de compreender dado que ele consiste em diversas árvores de decisão e o resultado predito pelo algoritmo é o resultado agregado de todas estas árvores (Murphy, 2012), desta maneira é um modelo apropriado para questões preditivas. Uma das vantagens deste algoritmo é que ele possibilita diminuir a variância e a instabilidade do algoritmo da árvore de decisão (Murphy, 2012).

O terceiro algoritmo a ser aplicado foi *Artificial Neural Networks* (ANN), sendo que sua escolha se deve a sua eficiência para lidar com linearidade e não linearidade presente nas variáveis que afetam a satisfação do consumidor (Larasati et al., 2012; Leong, 2015), por não precisar de pressupostos de linearidade, distribuição, homocedasticidade e por lidar bem com multicolinearidade entre as variáveis (Lee, Leong, Hew e Ooi, 2013; Lin, 2007b), por ser mais eficiente no processo de aprendizagem e predição do que as metodologias estatísticas tradicionais (West, Brockett e Golden, 1997; Gronholdt e Martensen, 2005), por ser amplamente utilizado pela literatura existente acerca de comportamento do

consumidor (Ngai, Xiu e Chau ,2009) e por ser uma das mais eficientes para lidar com comportamento do consumidor (Garver, 2002).

Desta maneira optou-se por utilizar a Regressão Logística por ser um modelo paramétrico e para obter maior interepretabilidade em relação aos resultados obtidos. Além disso, foram selecionados dois modelos com maior flexibilidade e que tendem a ser mais acurados, contudo possuem menor interpretabilidade: *Random Forest* e *Artificial Neural Networks*. O objetivo é comparar a performance destes dois modelos para prever a satisfação dos consumidores de *e-commerce* e analisar as relações que estes algoritmos encontram entre as variáveis com a satisfação do consumidor.

3.3.3. APLICAÇÃO DOS ALGORITMOS

Foi utilizado o software R para aplicação dos algoritmos utilizados, mais especificamente o pacote *Caret.* Nos casos de algoritmos não paramétricos (Random Forest e *Aritificial Neural Networks*) foram adotadas técnicas de *Hyperparameter Tuning* para selecionar os hiperparâmetros[11] destes modelos que maximizam o AUC[12].

No algoritmo de *Random Forest*, o hiperparâmetro otimizado foi o Número de variáveis para as quais é possível dividir cada nó[13], sendo que os valores possíveis para este hiperparâmetro vão de um até o número de variáveis presentes na base de dados. O valor deste hiperparâmetro que maximiza o AUC do modelo é 4.

No algoritmo *Artificial Neural Networks* os hiperparâmetros otimizados foram o número de neurônios presentes na *hidden layer* e o valor de *decay,* o qual é um hiperparâmetro de regularização do modelo para evitar *overfitting*. O valor de neurônios que otimizou a performance do algoritmo (medido em AUC) foi 11 e o valor de d*ecay* foi de 0.064.

---

[11] Os hiperparâmetros contêm os dados que controlam o próprio processo de treinamento do modelo e que maximizam as métricas de predição do modelo.

[12] Uma métrica utilizada para medir a capacidade preditiva do modelo, a qual será explicado adiante.

[13] Foi utilizada a função *ranger* do pacote *ranger* disponível para o *software* R, sendo que o nome do hiperparâmetro é *mtry*

A aplicação dos algoritmos e a análise de performance foram baseadas nas metodologias tradicionais de *Machine Learning* para classificação, descritas em Alpaydin (2010), Kelleher et al. (2015) e Flach (2012). Primeiramente as bases de dados foram divididas em dois conjuntos: treinamento e teste. No primeiro conjunto ficou uma amostra aleatória de 90% dos dados originais e no segundo 10% das observações dos dados originais. O conjunto de treinamento foi designado para treinar os algoritmos e o conjunto de teste foi utilizado para analisar a performance do algoritmo, avaliando qual o seu poder de previsão em um conjunto de dados que não foi utilizado para seu treinamento.

3.3.4. ANÁLISE DE PERFORMANCE E ROBUSTEZ DOS ALGORITMOS

Após o treinamento de cada algoritmo no conjunto de treinamento ele foi aplicado ao conjunto teste. Como as aplicações dos algoritmos têm o propósito de classificação foi gerada a Matriz de Confusão, que compara o *output* real do conjunto de treinamento com o que foi predito pelo algoritmo ao ser aplicado neste mesmo conjunto de dados. Na matriz de confusão a classe da variável dependente é Satisfação ou Não-Satisfação do consumidor, dessa maneira, os casos em que o algoritmo prevê corretamente que o consumidor está satisfeito são chamados de *True Positive* (Verdadeiro Positivo), os casos em que o algoritmo previu de maneira errada (isto é, o algoritmo assinala o consumidor como Não-Satisfeito quando na verdade ele está satisfeito) são denominados de *False Positive* (Falso Positivo). Os casos em que o algoritmo prevê corretamente que o consumidor não está satisfeito é chamado de *True Negative* (Verdadeiro Negativo) e os casos em que o algoritmo prevê errado a Não-Satisfação do consumidor é denominado de *False Negative* (Falso Negativo). Dessa maneira a matriz de confusão é construída, podendo ser visualizada a seguir:

QUADRO 2: MATRIZ DE CONFUSÃO

	Resultado previsto	
Resultado real	Satisfeito	Não-Satisfeito
Satisfeito	*True Positive*	*False Positive*
Não-Satisfeito	*False Negative*	*True Negative*

Fonte: Elaborado pelo autor com base em Alpaydin (2010), Kelleher et al. (2015) e Flach (2012)

A partir dos resultados de predição dos algoritmos foram obtidas quatro métricas para avaliar o desempenho do mesmo. Primeiramente foi calculada a acurácia, que é obtida a partir das medidas de *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) e *False Negative* (TN):

$$acurácia = \frac{TP+TN}{TP+TN+FP+TP} \quad (9)$$

A segunda métrica utilizada para avaliar a performance do algoritmo foi a precisão, que é calculada pela seguinte equação:

$$precisão = \frac{TP}{TP+FP} \quad (10)$$

A terceira métrica utilizada foi o *recall* que é definido da seguinte maneira:

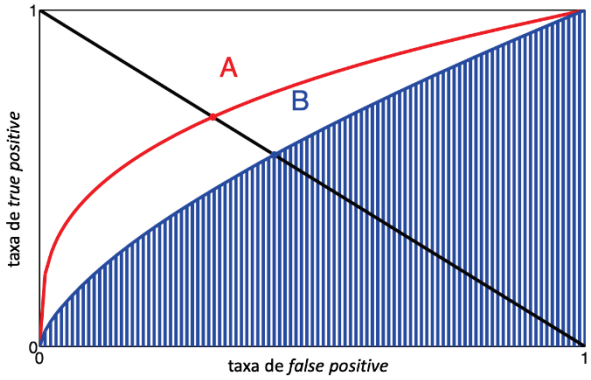
$$recall = \frac{TP}{TP+FN} \quad (11)$$

A última métrica utilizada foi a *AUC* (*Area Under the Curve*), isto é, a área abaixo da curva ROC (*receiver operating characteristics*) baseada nas taxas de *true positives* e *false positives*, sendo que AUC é a melhor medida para analisar a performance do algoritmo (Huang e Ling, 2005). ROC mostra o *trade-off* entre a taxa de *true positve* (que equivale ao *recall*) e a taxa de *false positive*, cuja equação é:

$$taxa\ de\ false\ positive = \frac{TN}{TP+FN} \quad (10)$$

Assim, a curva ROC mostra as diferentes combinações de taxa de *true positive* e *false positive* do algoritmo, sendo que a métrica AUC se refere a área abaixo dessa curva, como pode ser visto no gráfico a seguir:

FIGURA 2: CURVA RUC



Fonte: Retirado de Murphy (2014)

Observa-se na imagem que a melhor AUC é representada pela curva A, visto que ela possui uma maior taxa de *true positive* a uma menor taxa de *false positive*. O valor máximo de AUC é equivalente a 1 (quando a taxa de *true positive* é 1 e a taxa de *false negative* é 0. Contudo, em estudos relacionados à psicologia, o resultado de AUC maior ou igual a 0.7 já indica forte influência das variáveis na variável dependente (Rice e Harris, 2005).

Além disso, houve a preocupação de evitar problemas com *overfitting* dos três algoritmos, isto é, a situação em que o algoritmo se adapta exatamente com a estrutura dos dados do conjunto de treinamento de tal maneira que as predições para o conjunto teste não são confiáveis. Para melhor responder ao overfitting, foi utilizado a técnica de *K-fold Cross Validation* em que a base de dados é dividida randomicamente em *K* grupos iguais. Um destes K grupos selecionados é utilizado para ser o conjunto de validação e os demais K-1 são utilizados para treinamento do algoritmo. Isso é repetido utilizando cada um dos K grupos como conjunto de validação, para observar se existem diferenças significativas na estrutura dos algoritmos e em sua performance (cuja métrica utilizada foi o R-Quadrado) entre cada um dos K grupos (Alpaydin, 2010). Neste trabalho foi utilizado K = 10.

3.3.5. MODELAGEM PARAMÉTRICA DA DESCONFIRMAÇÃO

A partir da base de dados desenvolvida para a aplicação dos algoritmos, foram utilizados os campos *review*_score, *disconfirmation* e *payment_type,* além disso, foram trazidas as informações de estado do consumidor (*customer*_state) e categoria do produto comprado (*product_category*_name).

A forma funcional proposta por Kahneman e Tversky (1992) para a função valor é dada por:

$$v(x) = \begin{cases} x^{\alpha}, & se\ x \geq 0 \\ -\lambda(-x)^{\beta}, & se\ x < 0 \end{cases} \qquad (7)$$

Onde $v(x)$ é o valor atribuído pelo prospecto, *x* é o resultado a partir do ponto de referência (podendo ser positivo ou negativo), α é o parâmetro de curvatura para ganhos, β é parâmetro de curvatura para perdas e λ é o parâmetro de aversão a perda.

Esta mesma forma funcional da Função Valor foi utilizada para estimar os parâmetros α, β e λ. $v(x)$ se refere à nota dada pelo consumidor (*review_score*) e a variável *x* se refere à desconfirmação (positiva, caso maior que zero, e negativa, caso menor que zero).

Foi utilizado o método *Non Linear Least Squares* (NLLS)[14], possibilitando estimar os três parâmetros. Primeiramente foram estimados os parâmetros considerando a base de dados como um todo. Em seguida, os parâmetros foram estimados para 5 Estados brasileiros distintos: São Paulo (SP), Rio de Janeiro (RJ), Paraná (PR), Bahia (BH) e Minas Gerais (MG). Após estimar os parâmetros, foi utilizado o teste de hipótese *t-test* para analisar a significância estatística das diferenças entre os parâmetros de cada Estado. O mesmo foi feito para tipo de pagamento (boleto, cartão de crédito, cartão de débito e voucher) e para categoria de produto (*cama_mesa_banho*, *beleza_saude, esporte_lazer, informatica_acessorios* e *moveis_decoração*).

A partir dos três modelos desenvolvidos (Regressão Logística, *Random Forest* e *Artificial Neural Networks*) foi estimada a probabilidade de o consumidor

[14] Foi utilizado a função *nls* presente no pacote *Stats* da linguagem R.

estar satisfeito, sendo que tal probabilidade vai de 0 a 1. O valor desta probabilidade estimada pelos diferentes algoritmos para cada uma das observações foi utilizado como *proxy* para a satisfação do consumidor. A partir de tal probabilidade foi criada a variável Satisfação, através da multiplicação da probabilidade por 5 para se assemelhar à escala da variável *review_score,* fazendo com que a variável Satisfação vá de 0 (caso a probabilidade de estar satisfeito seja 0) até 5 (caso a probabilidade de estar satisfeito seja 1).

Desta maneira, utilizando *Non Linear Least Squares* (NLLS) foram estimados os três parâmetros α, β e λ a partir das probabilidades estimadas por cada um dos algoritmos, possibilitando a comparação dos três parâmetros entre os três algoritmos e em relação ao que foi estimado utilizando a variável *review_score*. Nestes casos, $v(x)$ se refere à Satisfação (calculada a partir da probabilidade calculada por cada um dos algoritmos para cada uma das observações de o consumidor estar satisfeito) e a variável *x* se refere à desconfirmação (positiva, caso maior que zero, e negativa, caso menor que zero).

3.3.6. INTERPRETAÇÃO DOS RESULTADOS

Como descrito por James et al. (2013), o resultado da Regressão Logística é altamente interpretável, sendo que a interpretação se dá justamente pelos parâmetros calculados pelo modelo. Contudo, devido ao *trade-off* entre interpretabilidade e a acurácia, apesar da Regressão Logística ser facilmente interpretável, ela tende a ser menos acuradas. Desta maneira, utilizou-se o algoritmo de *Random Forest* e de *Artificial Neural Networks,* os quais tendem a ser mais acurados, porém menos interpretáveis. Assim, para interpretar os resultados gerados por estes algoritmos foram utilizadas técnicas específicas para compreender quais as variáveis mais importantes para a predição e o impacto positivo ou negativo de cada variável independente do modelo.

A importância das variáveis para a predição do algoritmo foi calculada com base em Fisher, Rudin e Dominici (2018), sendo que a metodologia se baseia em calcular primeiramente o erro do modelo original

$$e^{orig} = L\ (y, f\ (X)) \qquad (2)$$

Onde y é o valor real do frete e f(x) é o valor predito pelo modelo. Em seguida, é gerada uma matriz de permutação (x^{perm}) para cada uma das variáveis e é calculado o erro do modelo baseado nas previsões realizadas pelo algoritmo com os dados permutados:

$$e^{perm} = L\ (Y, f\ (X^{perm})) \qquad (3)$$

Por fim, é calculada a importância para cada variável j através da seguinte fórmula:

$$FIj = e^{perm} / e^{orig} \qquad (4)$$

Por fim, para calcular a importância das variáveis em *Artificial Neural Networks*, foi utilizado o método de Olden (Olden et al., 2004). Ele calcula a importância das variáveis baseado da multiplicação dos pesos presentes nas conexões entre a *input-layer* e *hidden-layer* da rede neural com os pesos na conexão entre a *hidden-layer* e *output-layer*, permitindo avaliar tanto a magnitude quanto o sinal do impacto de cada variável no *output*.

A partir destas técnicas de interpretação de algoritmos é possível testar as hipóteses levantadas utilizando os algoritmos *Random Forest* e *Artificial Neural Networks.*

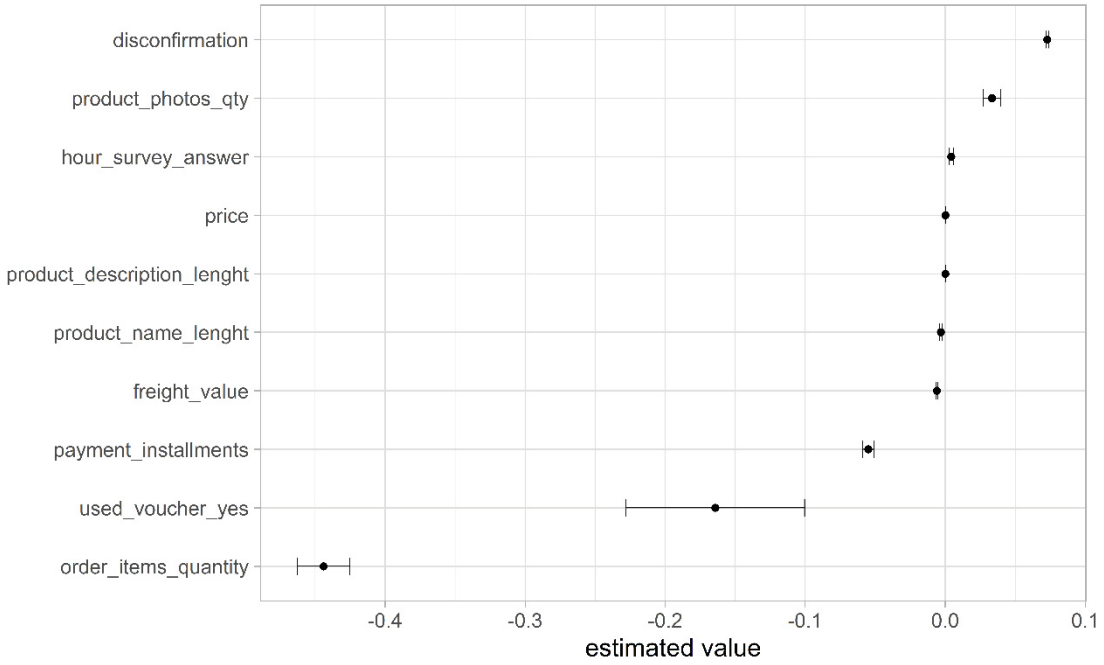
3.4. RESULTADOS

Os resultados do presente trabalho estão apresentados em duas seções distintas. Na primeira delas é apresentado o resultado dos três algoritmos (Regressão Logística, *Random Forest* e *Artificial Neural Networks*), com intuito de verificar se as hipóteses levantadas podem ser validadas ou não nos consumidores da Olist. Na segunda seção é apresentada uma análise específica para a relação da Desconfirmação com a satisfação do consumidor.

3.4.1. FATORES QUE AFETAM A SATISFAÇÃO DO CONSUMIDOR

O primeiro algoritmo aplicado foi a Regressão Logística, sendo este o único algoritmo paramétrico aplicado no presente trabalho e que possui maior interpretabilidade (James et al., 2013). O gráfico abaixo mostra os valores dos betas estimados e o desvio padrão (apresentados na Tabela X, na seção de anexos) para as variáveis que foram significativas:

GRÁFICO 20: BETAS ESTIMADOS E DESVIO PADRÃO DE VARIÁVEIS ESTATISTICAMENTE SIGNIFICATIVAS (p < 0.1) UTILIZANDO REGRESSÃO LOGÍSTICA



Fonte: Elaborado pelo autor

A variável com maior impacto positivo para a satisfação do consumidor foi a desconfirmação, com um beta de 0.0729. A variável desconfirmação indica justamente se a entrega atrasou em relação ao previsto (assumindo valores negativos), se foi exatamente igual ao previsto (assumindo valor de erro) ou se chegou antes do previsto (assumindo valores positivos). Neste caso, se observa que, conforme maior o valor da variável desconfirmação, maior a performance da entrega em relação à expectativa do consumidor, fazendo com que o consumidor fique mais satisfeito.

Este resultado está de acordo com a Hipótese 4, e acompanha trabalhos anteriores que identificaram um impacto positivo da desconfirmação na satisfação do consumidor (Mckinney et al., 2002; Susarla et al., 2006; Thong et al., 2006; Bhattacherjee, 2001; Park e Lee, 2011; Ho et al. 2017; . Szymanski e Henard, 2001). Além disso, este resultado também aponta para a Hipótese 2, isto é, que a Qualidade da Entrega é relevante para a satisfação do consumidor, conforme foi encontrado por Lin et al. (2011) e Ahn et al. (2004). Mais especificamente, o que este resultado mostra é que a Qualidade da Entrega possui um impacto positivo na satisfação do consumidor através do mecanismo de desconfirmação proposto por Mckinney et al (2002).

A segunda variável com maior impacto positivo na satisfação do consumidor foi a quantidade de fotos do produto, com beta estimado de 0.033. Este resultado aponta para a relevância da Qualidade da Informação na satisfação do consumidor, conforme encontrado por Lin (2007), Ahn et al. (2004) e DeLone e McLean (2003).

Em seguida, observa-se que a variável *payment_installments* (quantidade de parcelas) possui um impacto negativo na satisfação do consumidor, com um beta estimado de -0.05497. Isso indica que consumidores com maiores quantidades de parcelas tendem a ficar menos satisfeitos com o produto, evidenciando que a forma de pagamento possui relação com a satisfação do consumidor. Outra variável com impacto negativo relevante é a *used_voucher_yes*, que indica se o consumidor utilizou *voucher* na compra realizada, fazendo com que determinado valor seja abatido do valor da compra. O beta estimado para esta variável foi de -0.164, indicando que consumidores que utilizaram voucher em suas compras tendem a ficar menos satisfeitos com o produto ou serviço oferecido.

Por fim, a variável com maior impacto negativo foi a quantidade de itens comprados (*order_items_quantity*), com um beta estimado de -0.44, ou seja, conforme maior a quantidade de produtos comprados pelo consumidor menos satisfeito ele tende a ficar com a experiência de compra como um todo. As demais variáveis significativas do modelo tiverem um impacto próximo de zero.

No quadro abaixo é possível analisar a performance do algoritmo ao prever no conjunto teste:

QUADRO 3: MATRIZ DE CONFUSÃO DA REGRESSÃO LOGÍSTICA

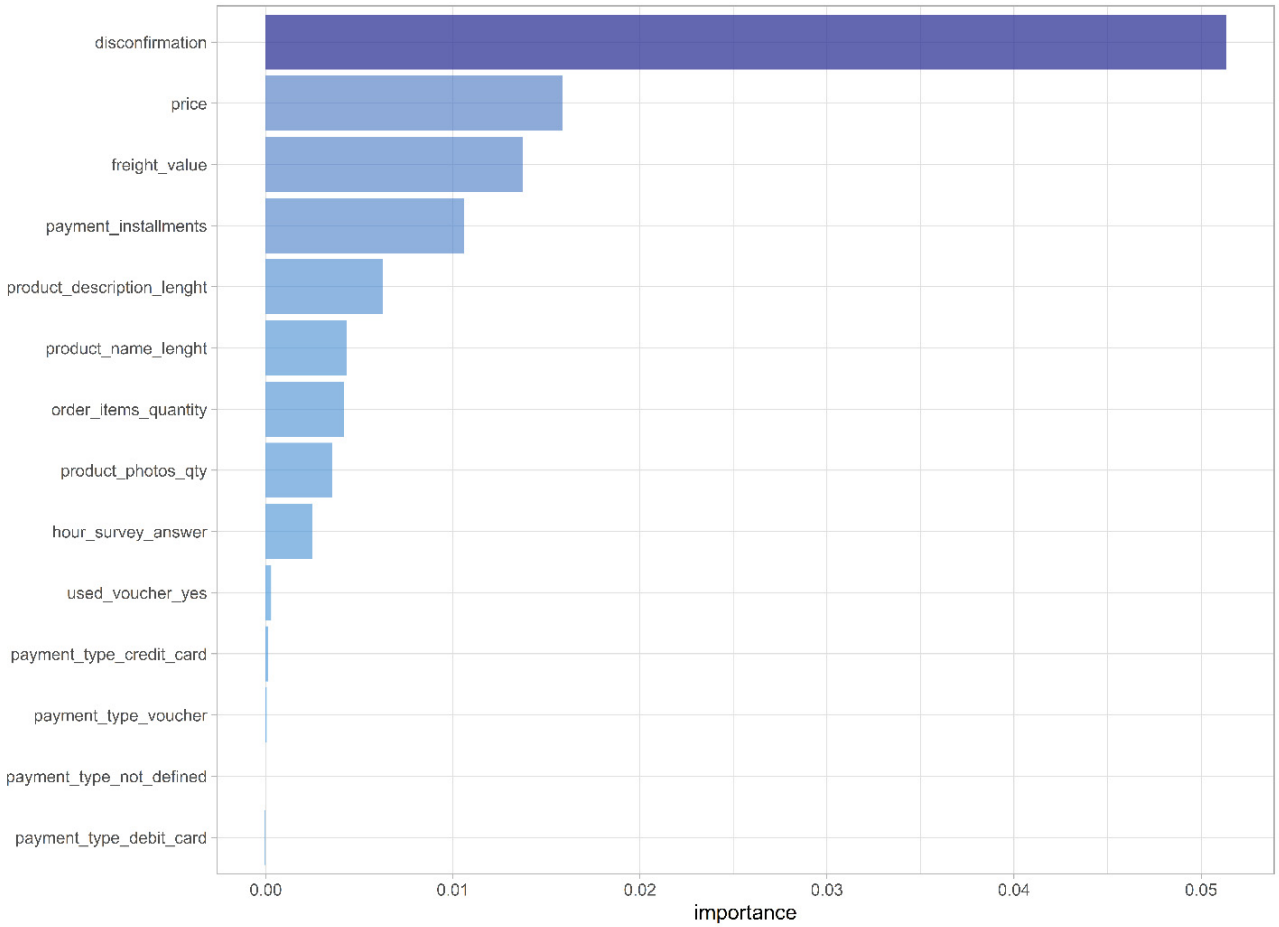
	Resultado previsto	
Resultado real	Satisfeito	Não-Satisfeito
Satisfeito	5615	2101
Não-Satisfeito	650	746

Fonte: Elaborado pelo autor

Observa-se que, quando o algoritmo prevê que o consumidor está satisfeito, ele tende a acertar, obtendo uma precisão de 89.6% e nos casos em que o consumidor realmente está satisfeito, o algoritmo tende a acertar 72.8%, sendo esta a métrica de *Recall.* De modo geral, o algoritmo se mostrou bom para acertar os casos em que o consumidor está satisfeito, contudo não foi eficiente para prever casos em que o consumidor está insatisfeito, sendo que a métrica de AUC foi 0.674, que é um valor relativamente baixo apesar de próxima do valor proposto por Rice e Harris (2005) de 0.7 para evidenciar impactos significativo no comportamento das pessoas.

O segundo modelo aplicado foi *Random Forest*, sendo que este modelo tende a ser mais acurado e menos interpretável (James et al., 2013). Desta maneira, para avaliar as hipóteses foi calculada a importância das variáveis a partir de permutações nas variáveis, conforme proposto por Fisher, Rudin e Dominici (2018). O gráfico abaixo mostra quais variáveis são mais importantes para prever a satisfação ou não do consumidor de acordo com o modelo:

GRÁFICO 21: IMPORTÂNCIA DAS VARIÁVEIS NA PREDIÇÃO DE SATISFAÇÃO DO CONSUMIDOR USANDO RANDOM FOREST



Fonte: Elaborado pelo autor

A variável mais importante para prever a satisfação do consumidor foi a desconfirmação, sendo que isto corrobora a Hipótese 4, isto é, que a desconfirmação exerce um papel importante na satisfação do consumidor, semelhante ao que foi encontrado por outros autores (Mckinney et al., 2002; Susarla et al., 2006; Thong et al., 2006; Bhattacherjee, 2001; Park e Lee, 2011; Ho et al. 2017; Szymanski e Henard, 2001). Contudo, não é possível afirmar se este impacto é positivo ou negativo a partir da importância das variáveis. Além disso, este resultado reforça novamente a Hipótese 2, isto é, a Qualidade da Entrega é relevante para a satisfação do consumidor (Lin et al., 2011; Ahn et al., 2004).

A segunda variável mais importante para prever a satisfação do consumidor foi o preço. Esse resultado aponta para a Hipótese 3, na qual o preço exerce um papel relevante na satisfação do consumidor (também encontrado por Cao e Gruca,

2004). Em seguida, se observa que o valor do frete é a terceira variável mais relevante para prever a satisfação do consumidor, seguido pela quantidade de parcelas (*payment_installments*). Por fim, as variáveis seguintes são variáveis relacionadas à Qualidade da Informação (tamanho da descrição do produto, tamanho do nome do produto e quantidade de fotos) juntamente com a quantidade de itens comprados. Este resultado mostra que a Qualidade da Informação também é relevante para a satisfação do consumidor, conforme apontado pela hipótese 1. As demais variáveis obtiveram uma importância baixa na predição da satisfação do consumidor.

A performance do algoritmo pode ser observada na matriz de confusão mostrada abaixo:

QUADRO 4: MATRIZ DE CONFUSÃO DE *RANDOM FOREST*

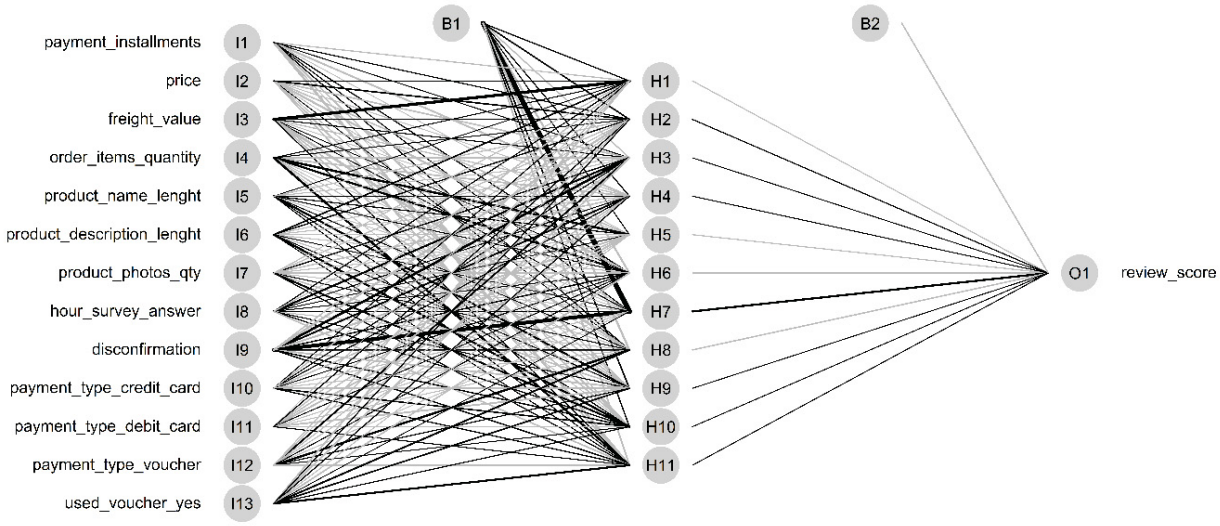
	Resultado previsto	
Resultado real	Satisfeito	Não-Satisfeito
Satisfeito	7053	663
Não-Satisfeito	740	656

Fonte: Elaborado pelo autor

O algoritmo de *Random Forest* obteve uma performance maior em todas as métricas avaliadas em relação à Regressão Logística, sendo que isto está de acordo com o que foi sustentado por James et al. (2013), evidenciando o trade-off entre interpretabilidade e acurácia. A acurácia do algoritmo em acertos foi de 84.6%, sendo que a sua precisão foi de 91.49% e seu *recall* foi de 91.73%. Por fim, o AUC do modelo foi de 0.749, sendo este valor acima do valor proposto por Rice e Harris (2005), indicando que as variáveis utilizadas pelo modelo possuem uma relação forte com o comportamento do consumidor.

O terceiro algoritmo aplicado foi *Artificial Neural Network.* A arquitetura da rede que otimizou o poder preditivo do modelo (medido em AUC) utilizou 11 neurônios na *hidden layer*, sendo que a estrutura da rede foi a primeira camada com os inputs (as 13 variáveis utilizadas no modelo), 11 neurônios na *hidden layer* e o *output* (o qual é binário, isto é, consumidor satisfeito ou insatisfeito), conforme pode ser observado na imagem abaixo:

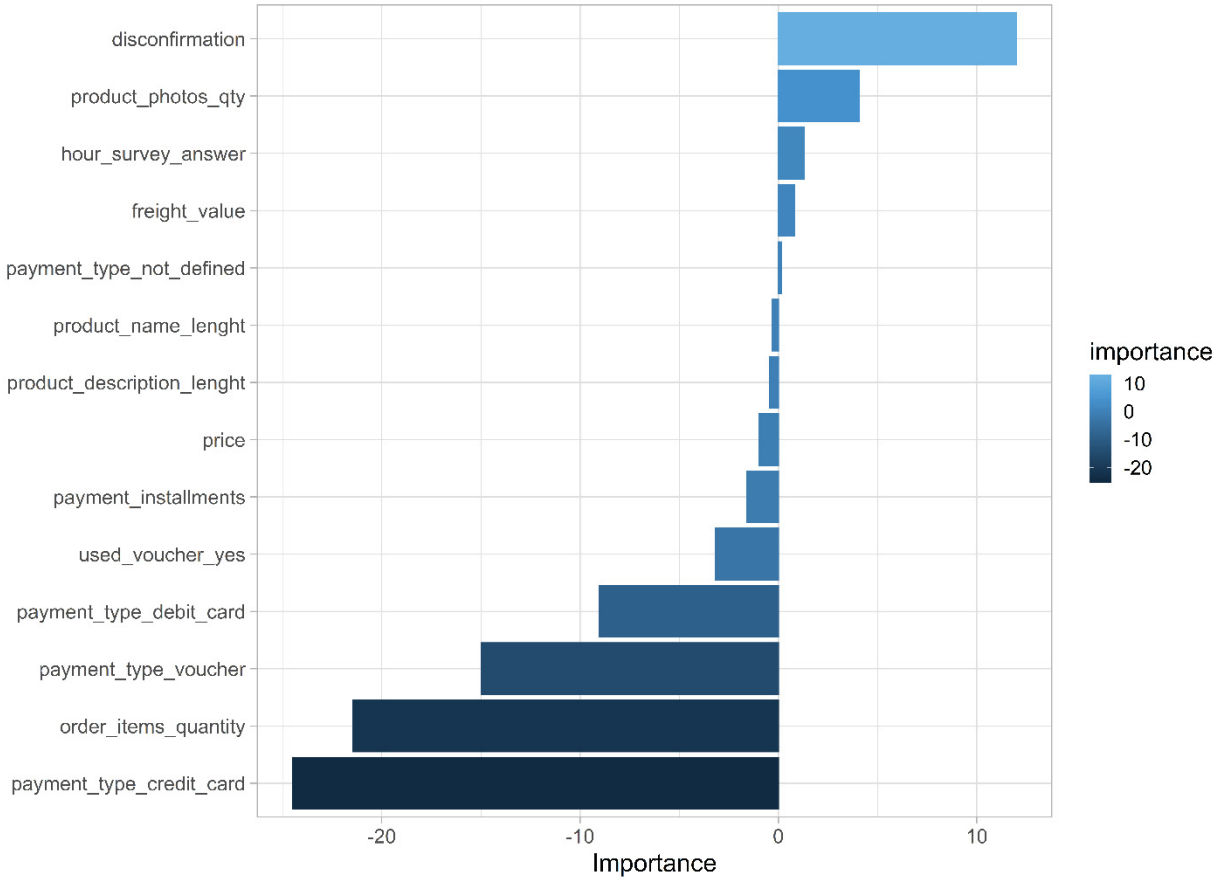
FIGURA 3: ARQUITETURA DE *ARTIFICIAL NEURAL NETWORKS* UTILIZADA E IMPACTOS ESTIMADOS



Fonte: Elaborado pelo autor

Observa-se de I1 até I13 todos os inputs, de H1 a H11 todos os neurônios e em O1 o *output*. Além disso, é possível notar os pesos que atualiza os pesos nos neurônios e no *output* durante o processo de *backpropagation* em B1 e B2, sendo que as linhas cinzas indicam impactos negativos e linhas pretas impactos positivos. Através da aplicação da metodologia de Olden (Olden et al., 2004) para medir a importância das variáveis em algoritmos de redes neurais, é possível compreender a magnitude e a direção do impacto de cada *input* na predição da satisfação ou insatisfação do consumidor, sendo que isto é representado no gráfico abaixo:

GRÁFICO 22: IMPORTÂNCIA DAS VARIÁVEIS NA PREDIÇÃO DE SATISFAÇÃO DO CONSUMIDOR USANDO A*RTIFICIAL NEURAL NETWORKS*



Fonte: Elaborado pelo autor

A variável com maior impacto positivo é a desconfirmação. Isso indica que conforme maiores os valores assumidos por esta variável (isto é, quando a performance é maior do que as expectativas), o consumidor tende a ficar mais satisfeito, semelhante ao que foi encontrado por outros autores (Mckinney et al., 2002; Susarla et al., 2006; Thong et al., 2006; Bhattacherjee, 2001; Park e Lee, 2011; Ho et al. 2017; . Szymanski e Henard, 2001). Esse resultado corrobora a Hipótese 4 de que a desconfirmação possui um impacto positivo e relevante na satisfação do consumidor.

A segunda variável com maior impacto positivo na satisfação do consumidor é a quantidade de fotos do produto, isto é, conforme a quantidade de fotos é maior, maior a probabilidade de o consumidor estar satisfeito. Este resultado evidencia novamente que a Qualidade da Informação é relevante para o consumidor ficar

satisfeito e que fotos são uma informação relevante para o consumidor, evidenciando a hipótese 1.

As variáveis com maiores impactos negativos são o uso de cartão de crédito pelo consumidor, a quantidade de itens comprados, pagamento com voucher, pagamento com débito e utilização de voucher. Isto indica que consumidores que utilizam meios de pagamento diferentes de boleto são os consumidores que tendem a ficar menos satisfeitos com o serviço ou produto oferecido pelo *e-commerce*. Além disso, o resultado está de acordo com o que foi mostrado pela Regressão Logística, em que a quantidade de itens afeta negativamente a satisfação do consumidor, isto é, consumidores que compram mais produtos em uma mesma compra tendem a ficar insatisfeitos.

A performance do algoritmo *ANN* pode ser observada abaixo na matriz de confusão:

QUADRO 5: MATRIZ DE CONFUSÃO DE *ARTIFICIAL NEURAL NETWORKS*

	Resultado previsto	
Resultado real	Satisfeito	Não-Satisfeito
Satisfeito	7642	74
Não-Satisfeito	1027	369

Fonte: Elaborado pelo autor

O algoritmo obteve uma acurácia de 87.9%, superando os dois algoritmos anteriores nesta métrica. Dos casos em que ele previu que o consumidor estaria satisfeito, ele acertou 86.1%, sendo esta sua métrica de Precisão, a qual foi inferior aos dois algoritmos anteriores. Por fim, dos casos que estão satisfeitos, o algoritmo acertou 99% deles, sendo esta sua métrica de *Recall.* Por fim, o AUC foi de 0.724, o qual apesar de ser pouco inferior ao do *Random Forest*, está acima de 0.70, indicando que as variáveis utilizadas pelo algoritmo exercem impacto no comportamento do consumidor (Rice e Harris, 2005). A tabela abaixo compara as 4 métricas para os 3 algoritmos utilizados:

TABELA 5: ALGORITMOS E SUAS MÉTRICAS DE PERFORMANCE

Métricas	Regressão Logística	Random Forest	ANN
Acurácia	0.698	0.846	0.879
Precisão	0.896	0.9149	0.861
Recall	0.728	0.9173	0.99
AUC	0.674	0.749	0.724

Fonte: Elaborado pelo autor

3.4.2. DESCONFIRMAÇÃO

As seções 4.1 e 4.2 anteriores buscaram analisar as hipóteses 1 a 5, procurando compreender os impactos das diferentes variáveis na satisfação do consumidor. Nesta seção serão abordadas as hipóteses 6 e 7, as quais buscam validar se o impacto da Desconfirmação na satisfação do consumidor segue a forma funcional da Função Valor proposta por Kahneman e Tversky (1992). Procura analisar se existe uma assimetria de impacto da desconfirmação na satisfação do consumidor, conforme proposto por Youjae Yi e Suna La (2003), em que a Desconfirmação Negativa tem um impacto maior na satisfação do que um mesmo montante de Desconfirmação Positiva. Esse fenômeno é evidenciado com a presença do parâmetro λ de aversão a perda da equação número. Além disso, busca-se avaliar se os valores destes parâmetros mudam conforme o Estado do consumidor, o tipo de pagamento e a categoria do produto.

Primeiramente foram estimados os parâmetros α, β e λ utilizando o método Non-Linear Least Square (NLLS), sendo que os valores estimados podem ser observados abaixo, todos estatisticamente significativos:

TABELA 6: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE

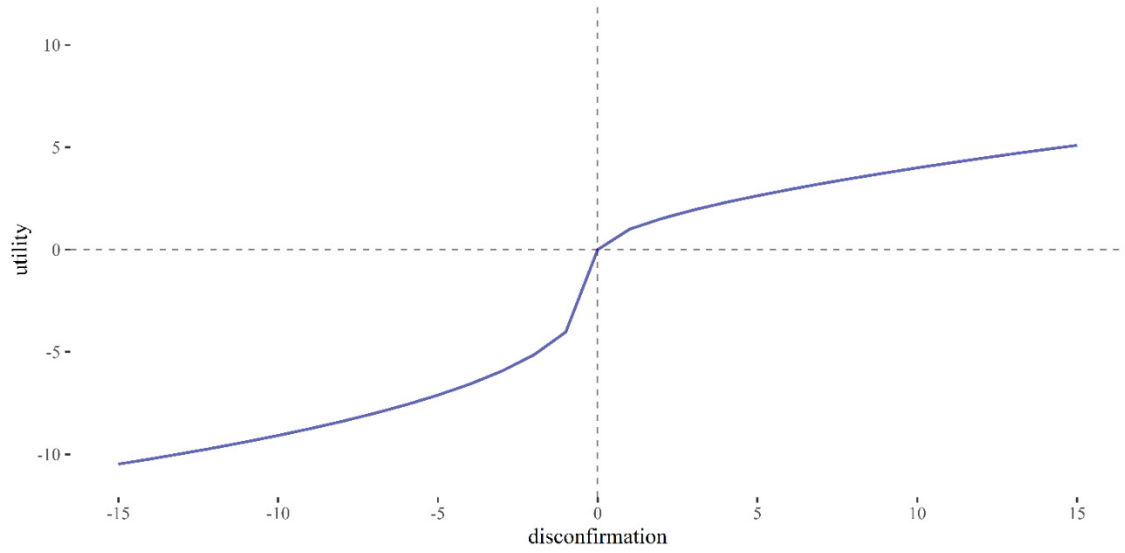
term	estimate	std error	p.value	significance	observations
α	0,601589	0,00070088	0	***	70721
β	0,353259	0,007482325	0	***	6736
λ	4,02576	0,036065715	0	***	6736

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; **.** p < 0.1

O primeiro ponto importante a ser observado é que os parâmetros α e β são menores do que 1, indicando a concavidade para ganhos e convexidade para perdas, conforme encontrado por outros autores (Tversky e Kahneman, 1992; Camerer e Ho, 1994; Wu e Gonzalez, 1996; Fennema e van Assen, 1998; Gonzalez e Wu, 1999; Abdellaoui, 2000; Donkers et al., 2001; Schmidt e Traub, 2002; Etchart-Vincent, 2004; Abdellaoui et al., 2005; Tu, 2005; Fehr-Duda et al., 2006; Abdellaoui et al., 2007; Andersen et al., 2006; Harrison and Rutström, 2009; Booij et al., 2010). Isto mostra justamente que há sensibilidade decrescente dos consumidores tanto em relação a perdas, quanto a ganhos, conforme teorizado por Kahneman e Tversky (1979; 1982).

Além disso, nota-se que o parâmetro β é menor do que o parâmetro α. Isto é diferente da maior parte do que a literatura encontrou ao estimar tais parâmetros. O que este resultado indica é que, no caso da desconfirmação, ganhos são avaliados mais ingremes que perdas, indicando que as pessoas se tornam menos sensíveis a perdas adicionais do que a ganhos adicionais (Booji et al., 2010). Por fim, o parâmetro λ equivale a 4, apontando para o que foi previsto por Tversky e Kahneman (1992), no qual o parâmetro λ deve ser maior que 1 para que haja presença de aversão a perda. Logo, todas as características da curva da Função Valor proposta por Tversky e Kahneman (1992), isto é, concavidade para ganhos, convexidade para perdas e aversão a perdas, foram identificadas na relação entre satisfação do consumidor com a Desconfirmação. Isto corrobora a Hipótese 5, indicando que o impacto da desconfirmação é assimétrico para ganhos e perdas, conforme proposto por Youjae Yi e Suna La (2003) e que obedece a forma funcional da Função Valor proposta por Tversky e Kahneman (1992). O gráfico abaixo mostra a forma da Função Valor utilizando os parâmetros estimados:

GRÁFICO 23: CURVA DE FUNÇÃO VALOR UTILIZANDO DESCONFIRMAÇÃO A PARTIR DOS PARÂMETROS ESTIMADOS



Fonte: Elaborado pelo autor

Em seguida, estes três parâmetros foram estimados para 4 formas distintas de pagamento: Cartão de crédito, boleto, cartão de débito e voucher. Todos os parâmetros foram estatisticamente significativos sendo que a tabela abaixo mostra a comparação dos parâmetros:

TABELA 7: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE POR MÉTODO DE PAGAMENTO

term	estimate	std.error	p.value	significance	observations	group
α	0,610580425	0,001641998	0	***	13173	boleto
β	0,308929842	0,015674731	5,15694E-76	***	1324	boleto
λ	4,290620772	0,083873663	0	***	1324	boleto
α	0,604709576	0,000820362	0	***	48287	credit_card
β	0,374491487	0,009005193	0	***	4596	credit_card
λ	4,0267353	0,042989104	0	***	4596	credit_card
α	0,622800652	0,006241432	0	***	1012	debit_card
β	0,353725474	0,059671408	5,16706E-08	***	95	debit_card
λ	4,040521807	0,232831154	6,44784E-31	***	95	debit_card
α	0,598931364	0,005895935	0	***	1019	voucher
β	0,328771441	0,090435354	0,00054936	***	67	voucher
λ	3,705512525	0,453682738	1,47053E-11	***	67	Voucher

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; **.** p < 0.1

Foi realizado um t-test para analisar se a diferença entre os parâmetros é estatisticamente significativa[15], e todos foram significativos com exceção do parâmetro λ para cartão de crédito e cartão de débito. Em relação ao parâmetro α foi observado que ele é maior para pagamentos com cartão de débito do que para pagamentos com voucher. Isso indica que pessoas que utilizam voucher se tornam menos sensíveis a ganhos adicionais de desconfirmação do que pessoas que utilizam cartão de débito, cartão de crédito e boleto.

Em relação ao parâmetro β, o maior valor foi para o método de pagamento de cartão de crédito e o menor valor foi para pagamentos com voucher. Esse resultado indica que pessoas que utilizam voucher se tornam menos sensíveis a perdas adicionais de desconfirmação do que pessoas que utilizam os demais métodos de pagamento. Por fim, o parâmetro de aversão a perda foi maior para quem utiliza boleto e menor para quem utiliza voucher. Este resultado mostra que existem diferenças de sensibilidade (para perdas e ganhos) e aversão a perda da Função Valor entre os diferentes métodos de pagamento.

[15] Tabela com resultados apresentada no anexo

Os três parâmetros também foram estimados para consumidores de 5 estados brasileiros distintos: SP, RJ, MG, RJ e BA. O resultado pode ser observado na tabela abaixo:

TABELA 8: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE POR ESTADO DO CONSUMIDOR

term	estimate	std.error	p.value	significance	observations	group
α	0,613602	0,001018	0	***	34462	SP
β	0,343151	0,01329	1,4E-128	***	2193	SP
λ	4,078055	0,057619	0	***	2193	SP
α	0,588599	0,002289	0	***	7578	RJ
β	0,370564	0,01817	6,02E-80	***	1241	RJ
λ	3,846497	0,093153	4,2E-235	***	1241	RJ
α	0,590475	0,001903	0	***	7989	MG
β	0,356768	0,026006	3,83E-37	***	563	MG
λ	4,002701	0,114636	8,7E-143	***	563	MG
α	0,592568	0,002904	0	***	3516	PR
β	0,306405	0,044171	4,37E-11	***	223	PR
λ	4,053962	0,21147	6,74E-49	***	223	PR
α	0,583395	0,004287	0	***	2106	BA
β	0,340589	0,031253	2,24E-24	***	400	BA
λ	4,126827	0,182225	1,48E-73	***	400	BA

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

Todas as diferenças entre os parâmetros foram estatisticamente significativas com exceção do parâmetro β para BA e SP. O maior valor para o parâmetro α foi de SP e o menor foi de BA, indicando que consumidores de BA se tornam menos sensíveis a ganhos conforme a desconfirmação aumenta do que consumidores dos demais estados. Quanto ao parâmetro β, o maior valor é do RJ e o menor valor é do PR, sendo que isso indica que consumidores do PR se tornam menos sensíveis a perdas de desconfirmação do que os demais estados. Por fim, o estado com maior aversão a perda foi BA e o estado com menor aversão a perda foi RJ.

Por fim, os três parâmetros foram calculados para 5 categorias de produtos diferentes: cama_mesa_banho, beleza_saude, esporte_lazer, informatica_acessorios e moveis_decoracao. O resultado pode ser observado na tabela abaixo:

TABELA 9: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE POR CATEGORIA DO PRODUTO

term	estimate	std.error	p.value	significance	observations	group
α	0,584202	0,002362	0	***	7250	cama_mesa_banho
β	0,356068	0,023541	7,25E-45	***	694	cama_mesa_banho
λ	4,012102	0,112539	8E-159	***	694	cama_mesa_banho
α	0,611961	0,002371	0	***	5917	beleza_saude
β	0,372403	0,023586	4,56E-48	***	681	beleza_saude
λ	4,232992	0,120648	1,2E-154	***	681	beleza_saude
α	0,609204	0,002411	0	***	5447	esporte_lazer
β	0,35786	0,028041	2,06E-32	***	493	esporte_lazer
λ	4,087251	0,145355	2,3E-104	***	493	esporte_lazer
α	0,591869	0,002773	0	***	4664	informatica_acessorios
β	0,3601	0,028057	2E-32	***	460	informatica_acessorios
λ	4,123197	0,133765	9E-114	***	460	informatica_acessorios
α	0,591395	0,002666	0	***	5209	moveis_decoracao
β	0,331036	0,026287	3,83E-32	***	554	moveis_decoracao
λ	3,921013	0,123266	6,4E-127	***	554	moveis_decoracao

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; **.** p < 0.1

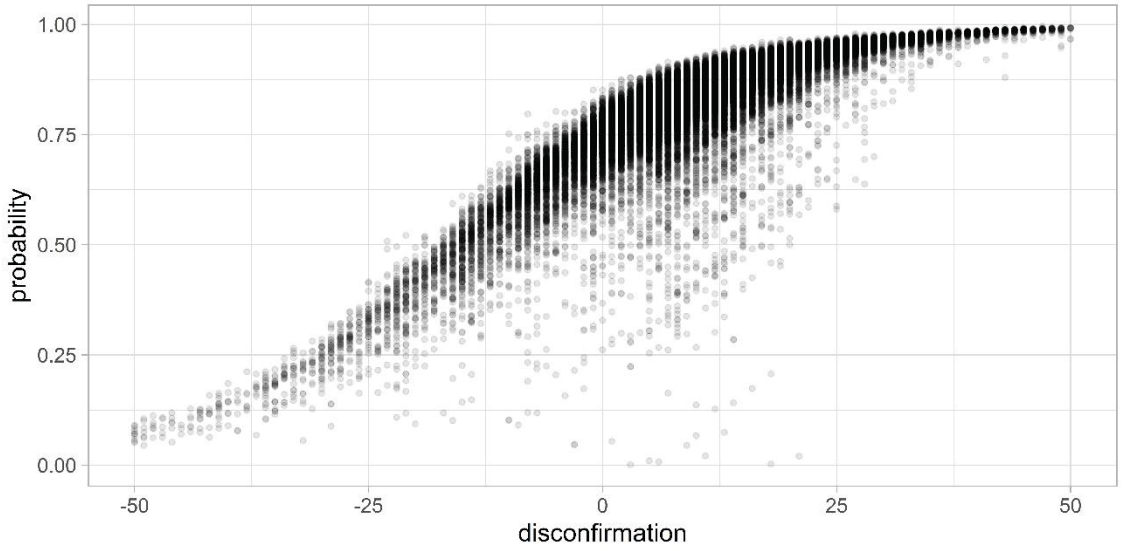
Todas as diferenças de valores foram estatisticamente significativas, com exceção do β para esporte_lazer e cama_mesa_banho, assim como, informática_acessorios e esporte_lazer. Em relação ao parâmetro α, o que se observa é que o maior valor é de beleza_saude e o menor valor é de cama_mesa_banho, sendo que isto indica que consumidores que compram produtos de cama_mesa_banho tendem a ter menor sensibilidade para ganhos do que consumidores de outras categorias de produto.

Em relação ao parâmetro β, foi observado que o maior valor foi para produtos de beleza_saude e o menor valor foi para moveis_decoracao, indicando que consumidores de moveis_decoracao se tornam menos sensíveis a perdas conforme suas perdas aumentam do que consumidores das demais categorias. Quanto ao parâmetro de aversão a perda, vemos que o maior valor foi para beleza_saude e o menor para moveis_decoracao.

Desta maneira, nota-se que existem diferenças entre os três parâmetros de acordo com o método de pagamento, estado do consumidor e categoria do produto, corroborando com o que é apresentado na Hipótese 6.

Em seguida, foi analisada a relação encontrada pelos três algoritmos utilizados (Regressão Logística, *Random Forest* e *Artificial Neural Networks*) entre a satisfação do consumidor com a desconfirmação. No gráfico abaixo é possível visualizar a relação entre a probabilidade de o consumidor estar satisfeito calculada pela Regressão Logística em relação a variável desconfirmação:

GRÁFICO 24: RELAÇÃO ENTRE PROBABILIDADE DE SATISFAÇÃO DO CONSUMIDOR E DESCONFIRMAÇÃO A PARTIR DA REGRESSÃO LOGÍSTICA



Fonte: Elaborado pelo autor

É possível observar que a relação entre a probabilidade estimada pelo algoritmo e a desconfirmação possuem justamente o formato da Função Valor proposta por Tversky e Kahneman (1979; 1992). Em seguida, foram estimados os parâmetros α, β e λ com intuito de analisar se a relação encontrada entre probabilidade de o consumidor estar satisfeito e desconfirmação seguem a forma funcional proposta por Kahneman e Tversky (1992). Na tabela abaixo é possível visualizar os três parâmetros estimados:

TABELA 10: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE A PARTIR PROBABILIDADE ESTIMADA POR REGRESSÂO LOGÍSTICA

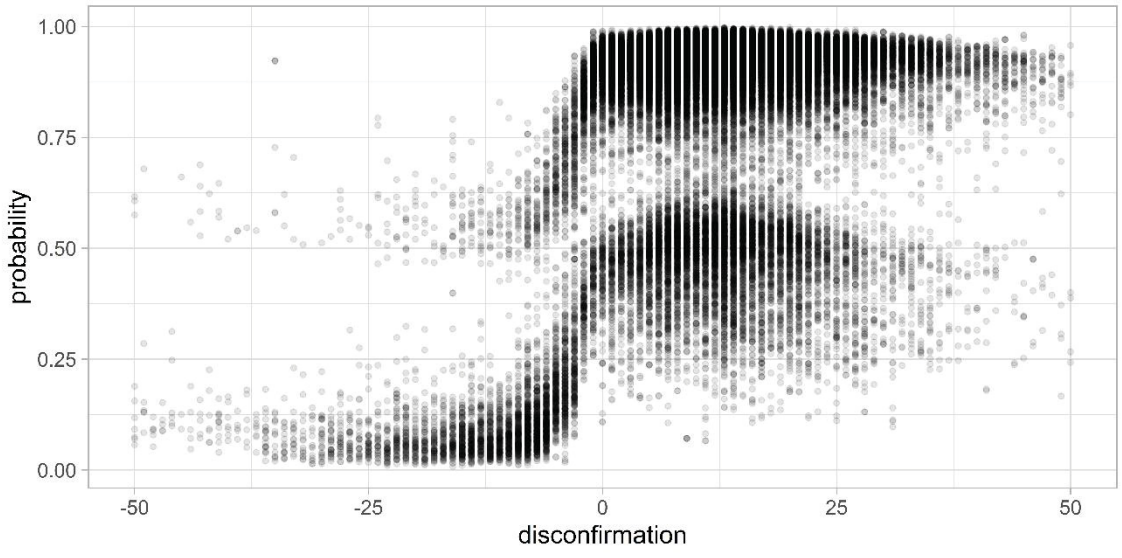
term	estimate	std.error	p.value	significance	n_obs
α	0,609109	0,000487	0	***	58102
β	0,108537	0,001729	0	***	5379
λ	3,745253	0,00998	0	***	5379

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

Todos os parâmetros foram estatisticamente significativos, sendo que eles seguem justamente o que foi teorizado por Kahneman e Tversky (1992), isto é, α é menor que um, β menor que um e λ maior que um, mostrando justamente a concavidade para ganhos, convexidade para perdas e a aversão a perda.

Este mesmo processo foi feito com os demais algoritmos, sendo que no gráfico abaixo é possível visualizar a relação entre a probabilidade estimada pelo algoritmo de *Random Forest* com a variável desconfirmação:

GRÁFICO 25: RELAÇÃO ENTRE PROBABILIDADE DE SATISFAÇÃO DO CONSUMIDOR E DESCONFIRMAÇÃO A PARTIR DE RANDOM FOREST



Fonte: Elaborado pelo autor

A partir dos dados de probabilidade da satisfação do consumidor estimada e da desconfirmação, foram estimados os três parâmetros para analisar a relação que

o algoritmo de *Random Forest* encontrou, sendo que na tabela abaixo é possível visualizar os resultados dos três parâmetros estimados:

TABELA 11: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE A PARTIR PROBABILIDADE ESTIMADA POR RANDOM FOREST

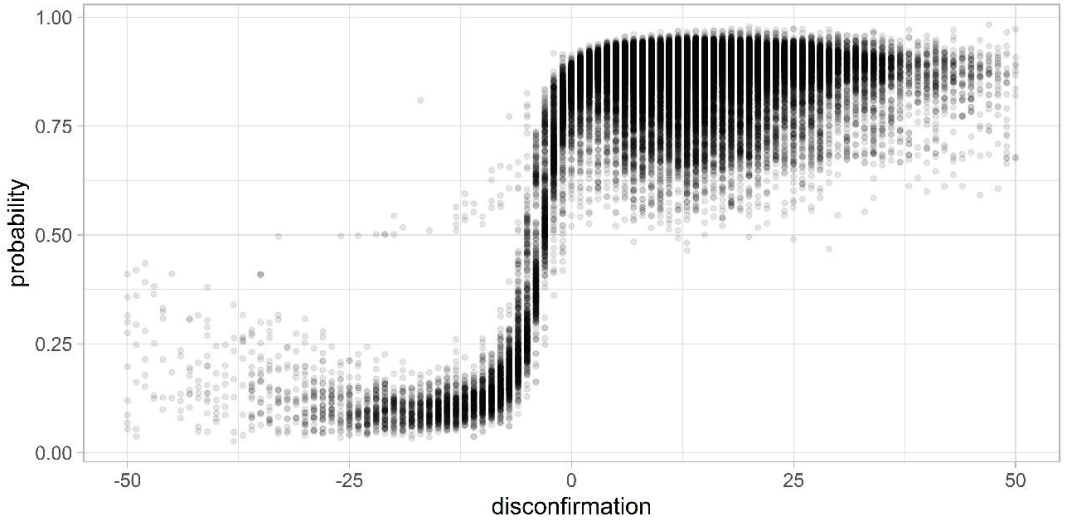
term	estimate	std.error	p.value	significance	n_obs
α	0,622275	0,00063	0	***	58102
β	0,639442	0,008481	0	***	5379
λ	4,329649	0,032024	0	***	5379

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; **.** p < 0.1

Semelhante ao que foi encontrado na Regressão Logística, todos os parâmetros apresentam o comportamento teorizado por Kahneman e Tversky (1992), isto é, convexidade para perdas, concavidade para ganhos e aversão a perda. A diferença de valores estimados em relação à regressão logística é que o segundo algoritmo capturou uma quantidade maior de aversão a perda do que primeiro (medido pelo parâmetro λ). Além disso, o valor encontrado para o parâmetro β foi maior que o valor encontrado na Regressão Logística e maior que o parâmetro α, indicando, sendo que este resultado é semelhante ao que é encontrado pela literatura (Booij et al., 2010).

Por fim, utilizando o algoritmo *Artificial Neural Networks*, foi analisada a relação entre a probabilidade estimada pelo algoritmo com a desconfirmação, sendo que o gráfico abaixo mostra esta relação:

GRÁFICO 26: RELAÇÃO ENTRE PROBABILIDADE DE SATISFAÇÃO DO CONSUMIDOR E DESCONFIRMAÇÃO A PARTIR DE *ARTIFICIAL NEURAL NETWORKS*



Fonte: Elaborado pelo autor

Em seguida, foram estimados os parâmetros que descrevem a relação entre a probabilidade estimada pelo algoritmo e a desconfirmação, sendo que todos eles foram estatisticamente significativos:

TABELA 12: PARÂMETROS ESTIMADOS UTILIZANDO NON-LINEAR LEAST SQUARE A PARTIR PROBABILIDADE ESTIMADA POR ARTIFICIAL NEURAL NETWORKS

term	estimate	std.error	statistic	p.value	significance	n_obs
α	0,622634	0,000563	1105,662	0	***	58102
β	0,607514	0,004679	-129,841	0	***	5379
λ	4,236014	0,017996	-235,386	0	***	5379

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

Novamente, nota-se que os valores estimados dos parâmetros satisfazem as condições de concavidade para ganhos, convexidade para perdas e aversão a perdas, conforma teorizado por Kahneman e Tversky (1979; 1992). Desta maneira, os três algoritmos encontram a uma relação na forma funcional da Função Valor

entre a satisfação do consumidor e a desconfirmação, evidenciando que a Desconfirmação exerce um impacto positivo na satisfação do consumidor (evidenciando a Hipótese 4) e segue a forma funcional da Função Valor proposta por Kahneman e Tversky (1982), apontando para a Hipótese 5.

3.5. CONCLUSÃO

O presente trabalho abordou diferentes aspectos da satisfação do consumidor, com intuito de entender os impactos das diferentes variáveis. As hipóteses analisadas foram:

- Hipótese 1: A Qualidade da Informação afeta a satisfação do consumidor (medido pela quantidade de caracteres no título, quantidade de caracteres na descrição e quantidade de fotos);
- Hipótese 2: A Qualidade da Entrega afeta a satisfação do consumidor (medido através da variável desconfirmação, a qual compara a o tempo de entrega estimado com o tempo de entrega real);
- Hipótese 3: O preço do produto afeta a satisfação do consumidor;
- Hipótese 4: A Desconfirmação possui um impacto positivo e relevante na satisfação do consumidor;
- Hipótese 5: A relação da Desconfirmação com a satisfação do consumidor segue a forma funcional da Função Valor proposta por Kahneman e Tvserky (1992);
- Hipótese 6: Os parâmetros de concavidade, convexidade e aversão à perda variam conforme o Estado do consumidor, método de pagamento e categoria de produto comprado;

Primeiramente foi analisado o impacto da Qualidade da Informação, Qualidade da Entrega, Preço do Produto e Desconfirmação na satisfação do consumidor (Hipóteses 1 a 4). Para isto foram utilizados três algoritmos (Regressão Logística, *Random Forest* e *Artificial Neural Networks*) para compreender o impacto destas variáveis na satisfação do consumidor.

A partir dos resultados dos três algoritmos, foi identificado que Desconfirmação (e consequentemente a Qualidade da Entrega) é a variável que

exerce o maior impacto positivo na satisfação do consumidor, que a Qualidade da Informação (principalmente através da quantidade de fotos) exerce um impacto positivo na satisfação do consumidor.

Além disso, foi analisado se o impacto da Desconfirmação na satisfação do consumidor é explicado pelos fenômenos de sensibilidade decrescente (para ganhos e perdas) e aversão a perda propostos na Teoria do Prospecto (Kahneman e Tversky, 1979; 1992), fazendo com que o impacto assimétrico da Desconfirmação, encontrados por Youjae Yi e Suna La (2003) e também encontrados no presente trabalho, sejam explicados pela Função Valor da Teoria do Prospecto e suas propriedades. Em seguida, foi analisado se os parâmetros estimados da Função Valor variam conforme o método de pagamento, Estado do consumidor e categoria do produto. Os parâmetros da Função Valor foram estimados utilizando Non-Linear Least Square (NLLS), tanto utilizando a nota dada pelo consumidor, quanto utilizando a probabilidade estimada pelos três algoritmos de o consumidor estar satisfeito, para desta maneira verificar se os três algoritmos encontraram uma relação entre satisfação e desconfirmação que é explicado pela Função Valor e que possui as propriedades teorizadas por Kahneman e Tversky (1979; 1992).

Foi identificado que a relação entre satisfação do consumidor e a desconfirmação respeita todas as propriedades da Função Valor proposta por Kahneman e Tversky (1979;1992), sendo que os parâmetros α e β são menores que 1 (indicando a sensibilidade descrescente) e o parâmetro λ é maior que 1 (indicando a aversão a perda). Este resultado explica o comportamento assimétrico da Desconfirmação encontrado por Youjae Yi e Suna La (2003), isto é, que a Desconfirmação Negativa tem maior impacto do que a Desconfirmação Positiva, pode ser explicado pela Teoria do Prospecto e apresenta todas as propriedades especificadas pela Função Valor. Além disso, foi identificado que os parâmetros mudam conforme o método de pagamento, Estado do consumidor e categoria do produto.

REFERÊNCIAS

ABDELLAOUI, Mohammed; VOSSMANN, Frank; WEBER, Martin. Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. **Management science**, v. 51, n. 9, p. 1384-1399, 2005.

ABDELLAOUI, Mohammed; BLEICHRODT, Han; PARASCHIV, Corina. Loss aversion under prospect theory: A parameter-free measurement. **Management Science**, v. 53, n. 10, p. 1659-1674, 2007.

ABDELLAOUI, Mohammed; BLEICHRODT, Han; L'HARIDON, Olivier. A tractable method to measure utility and loss aversion under prospect theory. **Journal of Risk and uncertainty**, v. 36, n. 3, p. 245, 2008.

ALPAYDIN, E. Introduction to machine learning/Ethem Alpaydin. 2010.

ANDERSEN, Steffen; HARRISON, Glenn W.; RUTSTRÖM, E. Elisabet. **Choice behavior, asset integration and natural reference points**. Working Paper 06, 2006.

ATKINSON, Elizabeth J.; THERNEAU, Terry M. An introduction to recursive partitioning using the RPART routines. Rochester: Mayo Foundation, 2000.

AZAR, Ofer H. Relative thinking and task performance: Does a larger fixed payment reduce the perceived magnitude of the pay-for-performance component. In: **Ben-Gurion University of the Negev Working Paper**. 2006.

AZAR, Ofer H. Relative thinking theory. **The Journal of Socio-Economics**, v. 36, n. 1, p. 1-14, 2007.

AZAR, Ofer H. The effect of relative thinking on firm strategy and market outcomes: A location differentiation model with endogenous transportation costs. **Journal of Economic Psychology**, v. 29, n. 5, p. 684-697, 2008.

AZAR, Ofer H. Does relative thinking exist in mixed compensation schemes. **Ben-Gurion University of Negev Working Paper**, 2009.

AZAR, Ofer H. Do people think about absolute or relative price differences when choosing between substitute goods?. **Journal of Economic Psychology**, v. 32, n. 3, p. 450-457, 2011.

AZAR, Ofer H. Do consumers make too much effort to save on cheap items and too little to save on expensive items? Experimental results and implications for business strategy. **American Behavioral Scientist**, v. 55, n. 8, p. 1077-1098, 2011.

AZAR, Ofer H. Does relative thinking exist in real-world situations? A field experiment with bagels and cream cheese. **Economic Inquiry**, v. 49, n. 2, p. 564-572, 2011.

AZAR, Ofer H. Relative thinking in consumer choice between differentiated goods and services and its implications for business strategy. **Judgment and Decision Making**, v. 6, n. 2, p. 176, 2011.

AZAR, Ofer H. Firm strategy and biased decision making: The price dispersion puzzle. **Applied Economics**, v. 45, n. 7, p. 901-910, 2013.

AZAR, Ofer H. Optimal strategy of multi-product retailers with relative thinking and reference prices. **International Journal of Industrial Organization**, v. 37, p. 130-140, 2014.

BISHOP, Christopher M. et al. Pattern recognition and machine learning (information science and statistics). 2006.

BEHARA, Ravi S.; FISHER, Warren W.; LEMMINK, Jos GAM. Modelling and evaluating service quality measurement using neural networks. **International journal of operations & production management**, v. 22, n. 10, p. 1162-1185, 2002.

BERRY, Michael J.; LINOFF, Gordon. **Data mining techniques: for marketing, sales, and customer support**. John Wiley & Sons, Inc., 1997.

BLUM, Adam. **Neural networks in C++: an object-oriented framework for building connectionist systems**. New York: Wiley, 1992.

BHATTACHERJEE, Anol. Understanding information systems continuance: an expectation-confirmation model. **MIS quarterly**, p. 351-370, 2001.

BOCQUÉHO, Géraldine; JACQUET, Florence; REYNAUD, Arnaud. Expected utility or prospect theory maximisers? Assessing farmers' risk behaviour from field-experiment data. **European Review of Agricultural Economics**, v. 41, n. 1, p. 135-172, 2013.

BOOIJ, Adam S.; VAN PRAAG, Bernard MS; VAN DE KUILEN, Gijs. A parametric analysis of prospect theory's functionals for the general population. **Theory and Decision**, v. 68, n. 1-2, p. 115-148, 2010.

BOUNSAYTHIP, Catherine; RINTA-RUNSALA, Esa. Overview of data mining for customer behavior modeling. **VTT Information Technology Research Report, Version**, v. 1, p. 1-53, 2001.

BREIMAN, Leo. **Classification and regression trees**. Routledge, 2017.

BRIDGES, Eileen; YIM, Chi Kin; BRIESCH, Richard A. A high-tech product market share model with customer expectations. **Marketing Science**, v. 14, n. 1, p. 61-81, 1995.

BROWN, Irwin; JAYAKODY, Ruwanga. B2C e-commerce success: A test and validation of a revised conceptual model. **The Electronic Journal Information Systems Evaluation**, v. 11, n. 3, p. 167-184, 2008.

BUCKINX, Wouter; VAN DEN POEL, Dirk. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. **European Journal of Operational Research**, v. 164, n. 1, p. 252-268, 2005.

BUSHONG, Benjamin; RABIN, Matthew; SCHWARTZSTEIN, Joshua. A model of relative thinking. **Unpublished manuscript, Harvard University, Cambridge, MA**, 2015.

CAMERER, Colin F.; HO, Teck-Hua. Violations of the betweenness axiom and nonlinearity in probability. **Journal of risk and uncertainty**, v. 8, n. 2, p. 167-196, 1994.

CAO, Yong; GRUCA, Thomas S. The influence of pre-and post-purchase service on prices in the online book market. **Journal of Interactive Marketing**, v. 18, n. 4, p. 51-62, 2004.

CHIANG, Wei-yu Kevin; ZHANG, Dongsong; ZHOU, Lina. Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression. **Decision Support Systems**, v. 41, n. 2, p. 514-531, 2006.

CHEUNG, Christy MK; LEE, Matthew KO. User satisfaction with an internet-based portal: An asymmetric and nonlinear approach. **Journal of the American Society for Information Science and Technology**, v. 60, n. 1, p. 111-122, 2009.

CHONG, Alain Yee-Loong. A two-staged SEM-neural network approach for understanding and predicting the determinants of m-commerce adoption. **Expert Systems with Applications**, v. 40, n. 4, p. 1240-1247, 2013.

DELONE, William H.; MCLEAN, Ephraim R. The DeLone and McLean model of information systems success: a ten-year update. **Journal of management information systems**, v. 19, n. 4, p. 9-30, 2003.

DENG, Weijaw. Using a revised importance–performance analysis approach: The case of Taiwanese hot springs tourism. **Tourism management**, v. 28, n. 5, p. 1274-1284, 2007.

DENG, Wei-Jaw; CHEN, Wen-Chin; PEI, Wen. Back-propagation neural network based importance–performance analysis for determining critical service attributes. **Expert Systems with Applications**, v. 34, n. 2, p. 1115-1125, 2008.

DONKERS, Bas; MELENBERG, Bertrand; VAN SOEST, Arthur. Estimating risk attitudes using lotteries: A large sample approach. **Journal of Risk and uncertainty**, v. 22, n. 2, p. 165-195, 2001.

ETCHART-VINCENT, Nathalie. Is probability weighting sensitive to the magnitude of consequences? An experimental investigation on losses. **Journal of Risk and Uncertainty**, v. 28, n. 3, p. 217-235, 2004.

FEHR-DUDA, Helga; DE GENNARO, Manuele; SCHUBERT, Renate. Gender, financial risk, and probability weights. **Theory and decision**, v. 60, n. 2-3, p. 283-313, 2006.

FENNEMA, Hein; VAN ASSEN, Marcel. Measuring the utility of losses by means of the tradeoff method. **Journal of risk and uncertainty**, v. 17, n. 3, p. 277-296, 1998.

FISHER, Aaron; RUDIN, Cynthia; DOMINICI, Francesca. All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. **arXiv preprint arXiv:1801.01489**, 2018.

FLACH, Peter. **Machine learning: the art and science of algorithms that make sense of data**. Cambridge University Press, 2012.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics, p. 1189-1232, 2001.

FRIEDMAN, Jerome H. et al. Predictive learning via rule ensembles. **The Annals of Applied Statistics**, v. 2, n. 3, p. 916-954, 2008.

FRISCH, Deborah. Reasons for framing effects. **Organizational behavior and human decision processes**, v. 54, n. 3, p. 399-429, 1993.

GALAR, Mikel et al. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. **Pattern Recognition**, v. 44, n. 8, p. 1761-1776, 2011.

GARVER, Michael S. Using data mining for customer satisfaction research. **Marketing Research**, v. 14, n. 1, p. 8, 2002.

GEDEON, Tamás D. Data mining of inputs: analysing magnitude and functional measures. **International Journal of Neural Systems**, v. 8, n. 02, p. 209-218, 1997.

GLOROT, Xavier; BORDES, Antoine; BENGIO, Yoshua. Deep sparse rectifier neural networks. In: **Proceedings of the fourteenth international conference on artificial intelligence and statistics**. 2011. p. 315-323.

GOLDSTEIN, Alex et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation**. Journal of Computational and Graphical Statistics**, v. 24, n. 1, p. 44-65, 2015.

GONZALEZ, Richard; WU, George. On the shape of the probability weighting function. **Cognitive psychology**, v. 38, n. 1, p. 129-166, 1999.

GRÖMPING, Ulrike. Variable importance assessment in regression: linear regression versus random forest. **The American Statistician**, v. 63, n. 4, p. 308-319, 2009.

GRNHOLDT, Lars; MARTENSEN, Anne. Analysing customer satisfaction data: a comparison of regression and artificial neural networks. **International Journal of Market Research**, v. 47, n. 2, p. 121-130, 2005.

HARRISON, Glenn W.; RUTSTRÖM, E. Elisabet. Expected utility theory and prospect theory: One wedding and a decent funeral. **Experimental economics**, v. 12, n. 2, p. 133, 2009.

HEATH, Chip; SOLL, Jack B. Mental budgeting and consumer decisions. **Journal of consumer research**, v. 23, n. 1, p. 40-52, 1996.

HIRSHMAN, Samel; POPE, Devin; SONG, Jihong. Mental Budgeting versus Relative Thinking. In: **AEA Papers and Proceedings**. 2018. p. 148-52.

HO, Yi-Chun; WU, Junjie; TAN, Yong. Disconfirmation effect on online rating behavior: A structural model. **Information Systems Research**, v. 28, n. 3, p. 626-642, 2017.

HUANG, Jin; LING, Charles X. Using AUC and accuracy in evaluating learning algorithms. **IEEE Transactions on knowledge and Data Engineering**, v. 17, n. 3, p. 299-310, 2005.

HUNG, Shin-Yuan; YEN, David C.; WANG, Hsiu-Yu. Applying data mining to telecom churn management. **Expert Systems with Applications**, v. 31, n. 3, p. 515-524, 2006.

JAMES, Gareth et al. **An introduction to statistical learning**. New York: springer, 2013.

KAHNEMAN, D., & TVERSKY, A. (1979). Prospect Theory: An Analysis of Decision under Risk. **Econometrica**, *47*(2), 263

KAMAKURA, Wagner A. et al. Assessing the service-profit chain. **Marketing science**, v. 21, n. 3, p. 294-317, 2002.

KARSOLIYA, Saurabh. Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. **International Journal of Engineering Trends and Technology**, v. 3, n. 6, p. 714-717, 2012.

KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. **Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies**. MIT Press, 2015.

KENGPOL, Athakorn; WANGANANON, Worrapon. The expert system for assessing customer satisfaction on fragrance notes: Using artificial neural networks. **Computers & Industrial Engineering**, v. 51, n. 4, p. 567-584, 2006.

KOH, Hian Chye; GERRY, Chan Kin Leong. Data mining and customer relationship marketing in the banking industry. **Singapore Management Review**, v. 24, n. 2, p. 1, 2002.

KŐSZEGI, Botond; RABIN, Matthew. A model of reference-dependent preferences. **The Quarterly Journal of Economics**, v. 121, n. 4, p. 1133-1165, 2006.

KŐSZEGI, Botond; RABIN, Matthew. Reference-dependent risk attitudes. **American Economic Review**, v. 97, n. 4, p. 1047-1073, 2007.

KŐSZEGI, Botond; RABIN, Matthew. Reference-dependent consumption plans. **American Economic Review**, v. 99, n. 3, p. 909-36, 2009.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. 2012. p. 1097-1105.

LARASATI, Aisyah; DEYONG, Camille; SLEVITCH, Lisa. The application of neural network and logistics regression models on predicting customer satisfaction in a student-operated restaurant. **Procedia-Social and Behavioral Sciences**, v. 65, p. 94-99, 2012.

LEE, Voon-Hsien et al. Knowledge management: a key determinant in advancing technological innovation?. **Journal of Knowledge Management**, v. 17, n. 6, p. 848-872, 2013.

LEONG, Lai-Ying et al. Predicting the determinants of the NFC-enabled mobile credit card acceptance: A neural networks approach. **Expert Systems with Applications**, v. 40, n. 14, p. 5604-5620, 2013.

LEONG, Lai-Ying et al. An SEM–artificial-neural-network analysis of the relationships between SERVPERF, customer satisfaction and loyalty among low-cost and full-service airline. **Expert Systems with Applications**, v. 42, n. 19, p. 6620-6634, 2015.

LIANG, Ting Peng; TURBAN, Efraim; ARONSON, Jay E**. Decision Support Systems and Intelligent Systems**. 2005

LIN, Hsiu-Fen. The impact of website quality dimensions on customer satisfaction in the B2C e-commerce context. **Total Quality Management and Business Excellence**, v. 18, n. 4, p. 363-378, 2007.

LIN, Wen-Bao. The exploration of customer satisfaction model from a comprehensive perspective. **Expert Systems with Applications**, v. 33, n. 1, p. 110-121, 2007.

LARIVIÈRE, Bart; VAN DEN POEL, Dirk. Predicting customer retention and profitability by using random forests and regression forests techniques. **Expert systems with Applications**, v. 29, n. 2, p. 472-484, 2005.

MATZLER, Kurt et al. The asymmetric relationship between attribute-level performance and overall customer satisfaction: a reconsideration of the importance–performance analysis. **Industrial marketing management**, v. 33, n. 4, p. 271-277, 2004.

MCKINNEY, Vicki; YOON, Kanghyun; ZAHEDI, Fatemeh “Mariam”. The measurement of web-customer satisfaction: An expectation and disconfirmation approach. **Information systems research**, v. 13, n. 3, p. 296-315, 2002.

MITTAL, Vikas; ROSS JR, William T.; BALDASARE, Patrick M. The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions. **The Journal of Marketing**, p. 33-47, 1998.

MOON, Philip; KEASEY, Kevin; DUXBURY, Darren. Mental accounting and decision making:: The relationship between relative and absolute savings. **Journal of Economic Behavior & Organization**, v. 38, n. 2, p. 145-153, 1999.

MOWEN, Maryanne M.; MOWEN, John C. An empirical examination of the biasing effects of framing on business decisions. **Decision Sciences**, v. 17, n. 4, p. 596-602, 1986.

MOZER, Michael C. et al. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. **IEEE Transactions on neural networks**, v. 11, n. 3, p. 690-696, 2000.

MURPHY, Kevin. Machine learning, a probabilistic perspective. 2012.

NGAI, Eric WT; XIU, Li; CHAU, Dorothy CK. Application of data mining techniques in customer relationship management: A literature review and classification. **Expert systems with applications**, v. 36, n. 2, p. 2592-2602, 2009

OLIVER, Richard L. Cognitive, affective, and attribute bases of the satisfaction response. **Journal of consumer research**, v. 20, n. 3, p. 418-430, 1993.

OLIVER, Richard L. **Satisfaction: A Behavioral Perspective on the Consumer: A Behavioral Perspective on the Consumer**. Routledge, 2014.

PARK, Bong-Won; LEE, Kun Chang. Exploring the value of purchasing online game items. **Computers in Human Behavior**, v. 27, n. 6, p. 2178-2185, 2011.

Murphy, Kevin. Machine learning, a probabilistic perspective. 2014.

RANYARD, Rob; ABDEL-NABI, Deborah. Mental accounting and the process of multiattribute choice. **Acta psychologica**, v. 84, n. 2, p. 161-177, 1993.

RICE, Marnie E.; HARRIS, Grant T. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. **Law and human behavior**, v. 29, n. 5, p. 615-620, 2005.

RONG-DA LIANG, Austin; CHEN, Houn-Gee. Is that deal worth my money? The effect of relative and referent thinking on starting price under different promotion programs using hotel coupons in online auctions. **Computers in Human Behavior**, v. 28, n. 2, p. 292-299, 2012.

RYU, Kisang; HAN, Heesup. Influence of physical environment on disconfirmation, customer satisfaction, and customer loyalty for first-time and repeat customers in upscale restaurants. 2010.

SAVAGE, Leonard J. **The foundations of statistics**. Courier Corporation, 1972.

SCHIFFERSTEIN, H. N. J.; KOLE, A. P. W.; MOJET, J. Asymmetry in the disconfirmation of expectations for natural yogurt. **Appetite**, v. 32, n. 3, p. 307-329, 1999.

SCHMIDT, Ulrich; TRAUB, Stefan. An experimental test of loss aversion. **Journal of risk and Uncertainty**, v. 25, n. 3, p. 233-249, 2002.

SMITH, Kate A.; WILLIS, Robert J.; BROOKS, Malcolm. An analysis of customer retention and insurance claim patterns using data mining: A case study. **Journal of the operational research society**, v. 51, n. 5, p. 532-541, 2000.

SPRENG, Richard A.; MACKENZIE, Scott B.; OLSHAVSKY, Richard W. A reexamination of the determinants of consumer satisfaction. **The Journal of Marketing**, p. 15-32, 1996.

ŠTRUMBELJ, Erik; KONONENKO, Igor. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, v. 41, n. 3, p. 647-665, 2014.

SUSARLA, Anjana; BARUA, Anitesh; WHINSTON, Andrew B. Understanding the ‘service’component of application service provision: an empirical analysis of satisfaction with ASP services. In: **Information Systems Outsourcing**. Springer, Berlin, Heidelberg, 2006. p. 481-521.

SZAJNA, Bernadette; SCAMELL, Richard W. The effects of information system user expectations on their performance and perceptions. **Mis Quarterly**, p. 493-516, 1993.

TEAS, R. Kenneth. Expectations, performance evaluation, and consumers' perceptions of quality. **The journal of marketing**, p. 18-34, 1993.

THALER, Richard. Toward a positive theory of consumer choice. **Journal of Economic Behavior & Organization**, v. 1, n. 1, p. 39-60, 1980.

THALER, Richard. Mental accounting and consumer choice. **Marketing science**, v. 4, n. 3, p. 199-214, 1985.

THALER, Richard H. Anomalies: Saving, fungibility, and mental accounts. **Journal of economic perspectives**, v. 4, n. 1, p. 193-205, 1990.

THALER, Richard H. Mental accounting matters. **Journal of Behavioral decision making**, v. 12, n. 3, p. 183-206, 1999.

THONG, James YL; HONG, Se-Joon; TAM, Kar Yan. The effects of post-adoption beliefs on the expectation-confirmation model for information technology continuance. **International Journal of Human-Computer Studies**, v. 64, n. 9, p. 799-810, 2006.

TSAUR, Sheng-Hshiung; CHIU, Yi-Chang; HUANG, Chung-Huei. Determinants of guest loyalty to international tourist hotels—a neural network approach. **Tourism Management**, v. 23, n. 4, p. 397-405, 2002.

TU, Qin et al. **Empirical analysis of time preferences and risk aversion**. Tilburg University, School of Economics and Management, 2005.

TVERSKY, Amos; KAHNEMAN, Daniel. The framing of decisions and the psychology of choice. **science**, v. 211, n. 4481, p. 453-458, 1981.

TVERSKY, Amos; KAHNEMAN, Daniel. Advances in prospect theory: Cumulative representation of uncertainty. **Journal of Risk and uncertainty**, v. 5, n. 4, p. 297-323, 1992.

WEST, Patricia M.; BROCKETT, Patrick L.; GOLDEN, Linda L. A comparative analysis of neural networks and statistical methods for predicting consumer choice. **Marketing Science**, v. 16, n. 4, p. 370-391, 1997.

WU, George; GONZALEZ, Richard. Curvature of the probability weighting function. **Management science**, v. 42, n. 12, p. 1676-1690, 1996.

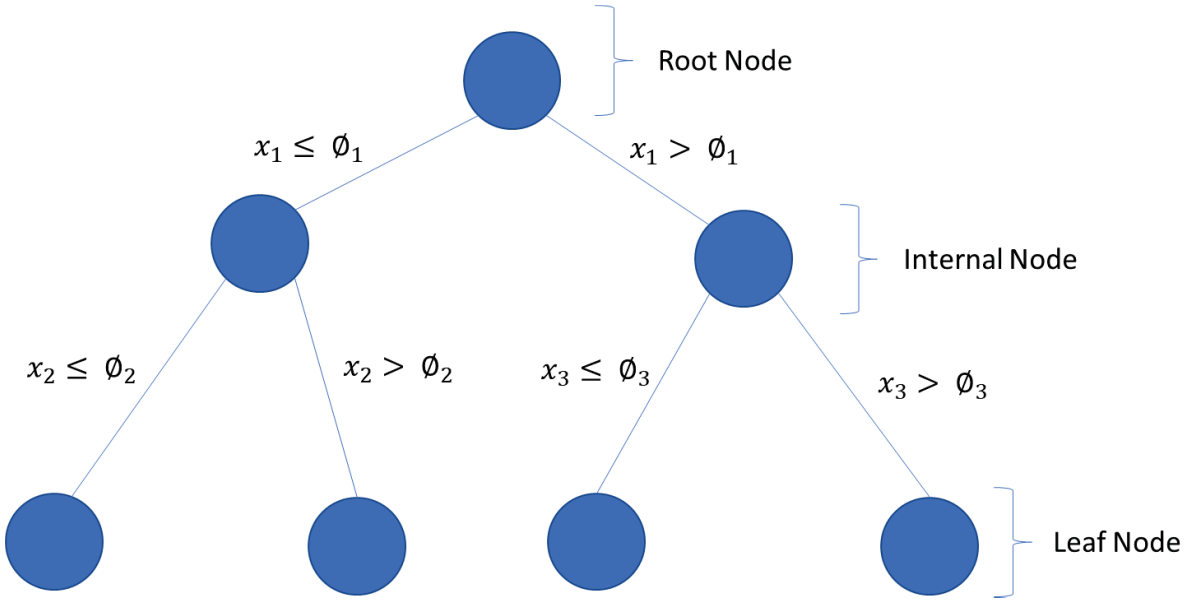
YI, Youjae; LA, Suna. The moderating role of confidence in expectations and the asymmetric influence of disconfirmation on customer satisfaction. **The Service Indsutries Journal**, v. 23, n. 5, p. 20-47, 2003.

ZEITHAML, Valarie A.; PARASURAMAN, Arun; MALHOTRA, Arvind. Service quality delivery through web sites: a critical review of extant knowledge. **Journal of the academy of marketing science**, v. 30, n. 4, p. 362, 2002.

APÊNDICE 1 – ÁRVORE DE DECISÃO

Baseado em Bishop (2006), o algoritmo inicia com o *root node* (nó raiz) que começa no topo da árvore, onde se encontra todas variáveis independentes (*inputs*). A partir deste primeiro nó o algoritmo começa um processo de particionamento binário recursivo no espaço total dos *inputs*. Primeiramente divide o espaço total dos *inputs* em duas regiões distintas de acordo com $x_1 \leq \theta_1$ ou $x_1 > \theta_1$ onde θ_1, é um parâmetro do modelo e x_1 uma das variáveis independentes, este processo cria duas sub-regiões distintas que são chamadas de *internal nodes* (nós internos), que serão subdivididas independentemente, por exemplo: a região onde $x_1 \leq \theta_1$ será dividida de acordo com um critério para a variável x_2, de acordo com $x_2 \leq \theta_2$ ou $x_2 > \theta_2$. Esse processo acontece recursivamente até o momento em que o *input* não é mais divisível de acordo com o critério utilizado, chegando assim ao último nó da árvore, denominado de *leaf node.* O processo pode ser observado na figura abaixo, baseada em Bishop (2006):

FIGURA 4: FUNCIONAMENTO DE ALGORITMO DE ÁRVORE DE DECISÃO



Fonte: Elaborado pelo autor com base em Bishop (2006)

Neste trabalho foi aplicado o algoritmo de *Decision Tree* para classificação binária (se o valor do frete está acima da média ou não). Foi utilizado o parâmetro de complexidade (que define quão específica será a árvore de decisão e o quanto ela

crescerá) de 0.01. O critério de decisão utilizado para dividir os nós e definir os parâmetros θ foi o Índice de Gini[16] mostrado por Breiman (2017). A equação que define o Índice de Gini (Bishop, 2006) pode ser visualizada abaixo:

$$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k}) \quad (11)$$

Onde, T representa o número total de *leaf nodes,* sendo que estes *leaf nodes* são indexados por $\tau = 1, \dots, |T|$, onde τ representa a região $\mathcal{R}_\tau$ do espaço total dos inputs, $p_{\tau k}$ é definido como a proporção dos dados na região $\mathcal{R}_\tau$ que pertencem à classe $\kappa = 1, \dots, K$, sendo que $Q_\tau(T)$ corresponde ao resíduo, sendo que o mesmo desaparece se $p_{\tau k} = 1$ ou $p_{\tau k} = 0$, sendo que o máximo é de $p_{\tau k} = 0.5$ (Bishop, 2006).

APÊNDICE 2 – *ARTIFICIAL NEURAL NETWORKS*

De acordo com Bishop (2006), ANN é composto por uma camada inicial chamado *input layer,* o qual se trata de todas as variáveis independentes do modelo, após esta camada vem uma segunda camada chamada de *hidden layer*, composta por neurônios, onde ocorre as interações entre as variáveis da primeira camada, sendo que pode haver mais de uma *hidden layer*, e por fim existe a camada de *output layer*, onde se encontram as variáveis dependentes.

Na camada inicial é construído M combinações lineares das variáveis dependentes ($x_1, \dots, x_D$) na seguinte forma:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \quad (12)$$

Onde a_j é denominada de ativação, $w_{ji}^{(1)}$ é o parâmetro que se refere ao peso (*weight*) para cada uma das variáveis independentes *i,* $w_{j0}^{(1)}$ é se refere ao viés, sendo que *j* = 1, ..., M e o termo (1) indicam o parâmetro correspondente dentro da primeira camada, isto é, se refere ao *j* neurônio da primeira *hidden layer.*

---

[16] A escolha deste critério se deve ao fato de ele ser melhor do que outros critérios como a taxa de classificação errada (*misclassification rate*) e ser mais sensitivo às probabilidades em cada um dos nós, além de ser um método de otimização baseados em gradientes e ser diferenciável (Bishop, 2006)

Cada uma das a_j são transformadas a partir de uma função de ativação *h(·)*, a qual é diferenciável e não linear, gerando então:

$$z_j = h(a_j) \tag{13}$$

Onde z_j é chamado de *hidden units*. A partir de cada z_j contido na *hidden layer* é combinado linearmente *K* vezes para cada um dos *K outputs*:

$$a_k = \sum_{i=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{14}$$

Estas transformações acontecem na segunda camada, que neste caso se refere ao *output layer* e a_k é denominada de unidade de ativação do *output*. Cada a_k é transformado através de uma função de ativação, gerando então o output y_k previsto pelo algortitmo:

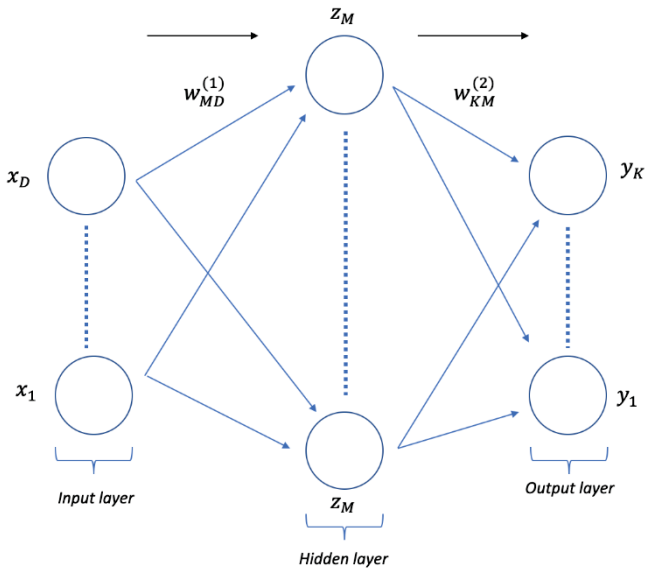
$$y_k = \sigma(a_k) \tag{15}$$

Sendo σ uma função de ativação genética, podendo ser $\sigma = h(\cdot)$ ou não. Agrupando as equações, temos que o ouput da ANN é dado por:

$$y_k = \sigma\left(\sum_{i=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{16}$$

A representação de todo o processo pode ser visualizada abaixo:

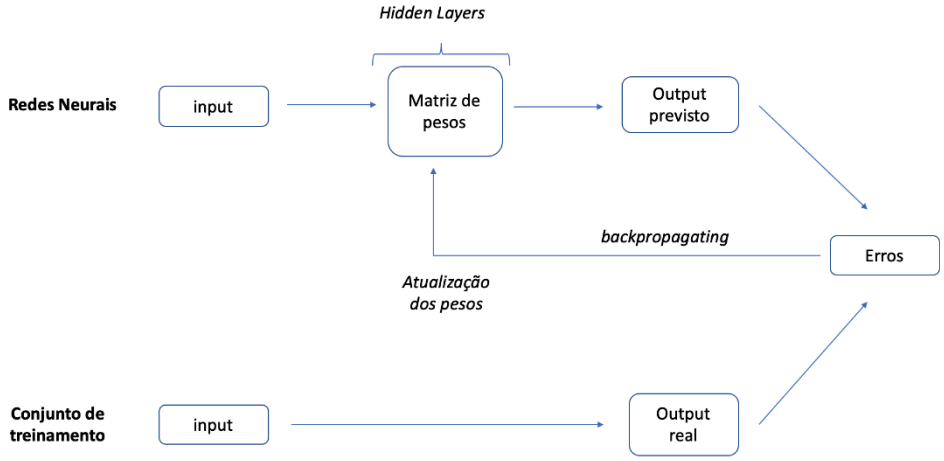
FIGURA 5: REPRESENTAÇÃO DO ALGORITMO ANN



Fonte: Elaborado pelo autor com base em Bishop (2006)

Medsker e Liebowitz (1994) recomendam o uso o uso de ANN com arquiterura de *multi-layer perceptron* para aplicações de classificação. Desta maneira, neste trabalho foi utilizado algoritmo ANN com *multi-layer perceptron* e com *Foward-Back Propagation*. O processo *Foward-Back Propagation* se inicia nos inputs que geram sinais que se movem pelas camadas de neurônios (*hidden layers*) em direção a variável dependente, ao longo desse processo é gerado uma matriz de pesos, descrito pela equação 2 (que mostra o impacto de cada neurônio na próxima camada), onde tais pesos são gerados a partir de uma função de ativação individual para cada neurônio, descrito pela equação 3. A partir deste processo e dos respectivos pesos, o algoritmo irá gerar os outputs previstos, como descritos nas equações 4 e 5, o qual é comparado com o output real do conjunto de treinamento, a partir desta comparação é gerado o erro não explicado do modelo, Este erro não explicado do modelo se propaga na direção contrária e esse processo se replica determinado número de vezes (sendo essa quantidade denominada de *epochs*), atualizando os parâmetros de peso de cada um dos neurônios ao longo do caminho. Esse processo é explicado em Liang et al. (2001) e Hagen et. Al (1996), sendo a útil a representação abaixo baseada nestes autores:

FIGURA 6: PROCESSO DE *FOWARD-BACK PROPAGATION*



Fonte: Elaborado pelo autor com base em Liang et al. (2001) e Hagen et. Al (1996)

ANEXO 1 – RESULTADOS DA REGRESSÃO LINEAR MÚLTIPLA

term	estimate	std.error	statistic	p.value	significance
(Intercept)	14,502	0,034973	414,6607	0	***
Price	0,484	0,037237	13,00342	1,76E-38	***
product_category_namedvds_blu_ray	0,177	0,037251	4,745295	2,10E-06	***
product_category_namerelogios_presentes	-0,238	0,061721	-3,85119	0,000118	***
product_category_nametelefonia	-0,334	0,088988	-3,75698	0,000173	***
customer_stateAL	0,481	0,09585	5,016785	5,31E-07	***
customer_stateMA	0,685	0,12559	5,456278	4,93E-08	***
customer_statePA	0,547	0,137383	3,97943	6,94E-05	***
customer_statePB	0,572	0,100234	5,703787	1,19E-08	***
customer_statePI	0,629	0,104401	6,029213	1,68E-09	***
customer_stateRN	0,375	0,095646	3,915961	9,04E-05	***
customer_stateRO	0,474	0,081854	5,787922	7,25E-09	***
customer_stateSE	0,443	0,087483	5,068537	4,05E-07	***
customer_stateTO	0,549	0,090462	6,068363	1,32E-09	***
distance	3,517	0,070661	49,77488	0	***
product_category_nameaudio	-0,118	0,040686	-2,90726	0,003651	**
product_category_namebrinquedos	-0,173	0,055126	-3,13734	0,001708	**
product_category_namepapelaria	-0,128	0,045183	-2,83746	0,004553	**
product_category_nametelefonia_fixa	-0,105	0,039848	-2,62364	0,008707	**
order_items_quantity	-0,07	0,03555	-1,96557	0,049364	*

product_category_namecasa_construcao	0,088	0,035829	2,456602	0,014035	*
product_category_nameconstrucao_ferramentas_ferramentas	0,084	0,03634	2,315435	0,020601	*
product_category_namecool_stuff	-0,104	0,047378	-2,19147	0,028431	*
product_category_nameesporte_lazer	-0,192	0,075597	-2,54604	0,010904	*
product_category_namemarket_place	-0,083	0,039367	-2,10978	0,034892	*
product_category_nameperfumaria	-0,154	0,061847	-2,49402	0,01264	*
customer_stateAM	0,149	0,073214	2,039135	0,041451	*
customer_stateDF	0,417	0,210491	1,980958	0,047612	*
customer_stateGO	0,483	0,203444	2,374855	0,017567	*
customer_stateMG	0,909	0,460479	1,973867	0,048412	*
customer_stateSC	0,686	0,274034	2,504176	0,012283	*
weight_used	-0,086	0,037711	-2,28668	0,022227	*
product_category_nameartes	0,064	0,035667	1,793531	0,072905	.
product_category_namebebes	-0,082	0,045354	-1,80896	0,070474	.
product_category_namefashion_roupa_infanto_juvenil	-0,067	0,035168	-1,90965	0,056195	.
product_category_namelivros_interesse_geral	0,064	0,036281	1,76571	0,077462	.
customer_stateMT	0,262	0,136441	1,916991	0,055255	.

customer_statePE	0,332	0,173596	1,910035	0,056145	.
customer_statePR	0,636	0,325999	1,950964	0,051077	.
customer_stateRJ	0,819	0,460639	1,77738	0,075523	.
product_category_namealimentos	-0,043	0,04242	-1,01452	0,310349	
product_category_namealimentos_bebidas	0,023	0,036027	0,640416	0,52191	
product_category_nameartes_e_artesanato	-0,017	0,03514	-0,49273	0,622211	
product_category_nameartigos_de_festas	-0,006	0,035435	-0,16783	0,866722	
product_category_nameartigos_de_natal	-0,015	0,036245	-0,41867	0,67546	
product_category_nameautomotivo	-0,066	0,066832	-0,98896	0,322697	
product_category_namebebidas	0,019	0,040087	0,462936	0,643416	
product_category_namebeleza_saude	-0,135	0,091892	-1,47436	0,140403	
product_category_namecama_mesa_banho	-0,018	0,04564	-0,40122	0,688265	
product_category_namecasa_conforto	-0,051	0,035358	-1,45336	0,146143	
product_category_namecine_foto	-0,021	0,036126	-0,57902	0,562581	
product_category_nameclimatizacao	0,018	0,03648	0,499786	0,617232	
product_category_nameconsoles_games	-0,066	0,055354	-1,20086	0,229821	
product_category_nameconstrucao_ferramentas_construcao	0,011	0,039191	0,276895	0,781864	
product_category_nameconstrucao_ferramentas_iluminacao	-0,036	0,036599	-0,99027	0,322054	
product_category_nameconstrucao_ferramentas_jardim	0,031	0,036597	0,855842	0,392097	
product_category_nameconstrucao_ferramentas_seguranca	0,006	0,036724	0,159455	0,873312	
product_category_nameeletrodomesticos	0,03	0,045236	0,660115	0,509189	
product_category_nameeletronicos	-0,093	0,08201	-1,12988	0,258541	
product_category_nameeletroportateis	-0,046	0,036627	-1,25216	0,210527	
product_category_namefashion_bolsas_e_acessorios	-0,12	0,074912	-1,60137	0,109312	
product_category_namefashion_calcados	-0,047	0,036738	-1,26804	0,204802	
product_category_namefashion_esporte	-0,02	0,035569	-0,56881	0,569494	
product_category_namefashion_roupa_feminina	-0,001	0,0351	-0,02035	0,983768	
product_category_namefashion_roupa_masculina	0,005	0,035227	0,139402	0,889134	
product_category_namefashion_underwear_e_moda_praia	0,008	0,038289	0,200371	0,841193	
product_category_nameferramentas_jardim	-0,03	0,040476	-0,75343	0,451201	
product_category_nameflores	-0,012	0,035083	-0,33267	0,739388	
product_category_namefraldas_higiene	0,005	0,035092	0,136361	0,891538	
product_category_nameindustria_comercio_e_negocios	-0,035	0,035052	-0,98541	0,324437	
product_category_nameinformatica_acessorios	-0,012	0,096	-0,12278	0,902285	
product_category_nameinstrumentos_musicais	-0,018	0,037639	-0,47861	0,63222	
product_category_namelivros_tecnicos	-0,013	0,035288	-0,37515	0,707555	
product_category_namemalas_acessorios	-0,002	0,036249	-0,04732	0,962255	

product_category_namemoveis_cozinha_area_de_servico_jantar_e_jardim	0,012	0,035089	0,354778	0,72276
product_category_namemoveis_decoracao	0,007	0,045993	0,155541	0,876396
product_category_namemoveis_quarto	-0,056	0,03565	-1,58451	0,113095
product_category_namemusica	-0,006	0,035768	-0,17271	0,86288
product_category_namepcs	0,021	0,035044	0,60971	0,542062
product_category_namepet_shop	0	0,047774	0,006304	0,99497
product_category_nameportateis_casa_forno_e_cafe	0,011	0,035379	0,316381	0,751717
product_category_namesinalizacao_e_seguranca	0,02	0,035739	0,567641	0,570286
product_category_nametablets_impressao_imagem	-0,042	0,035391	-1,20049	0,229965
product_name_lenght	0,054	0,037544	1,43055	0,152577
product_description_lenght	0,024	0,038993	0,60762	0,543447
product_photos_qty	0,053	0,039375	1,343786	0,179035
customer_stateAP	0,044	0,049634	0,890405	0,373261
customer_stateBA	0,315	0,253472	1,241248	0,214531
customer_stateCE	0,268	0,16905	1,583011	0,113437
customer_stateES	0,293	0,209286	1,399184	0,161776
customer_stateMS	0,057	0,125483	0,457885	0,647041
customer_stateRR	0,041	0,047191	0,863385	0,387938
customer_stateRS	0,412	0,317834	1,296885	0,194688
customer_stateSP	0,029	0,71911	0,039919	0,968158
trim	-0,052	0,035724	-1,44859	0,147471
hour	0,003	0,035213	0,071945	0,942647
day	-0,005	0,035154	-0,14528	0,884494
`weekdayQuarta Feira`	0,022	0,049008	0,458789	0,646391
`weekdayQuinta Feira`	-0,069	0,048334	-1,41753	0,156344
weekdaySábado	0,02	0,045473	0,444973	0,656345
`weekdaySegunda Feira`	-0,003	0,049374	-0,06548	0,947794
`weekdaySexta Feira`	-0,069	0,048073	-1,43099	0,15245
`weekdayTerça Feira`	-0,062	0,04936	-1,25428	0,209756

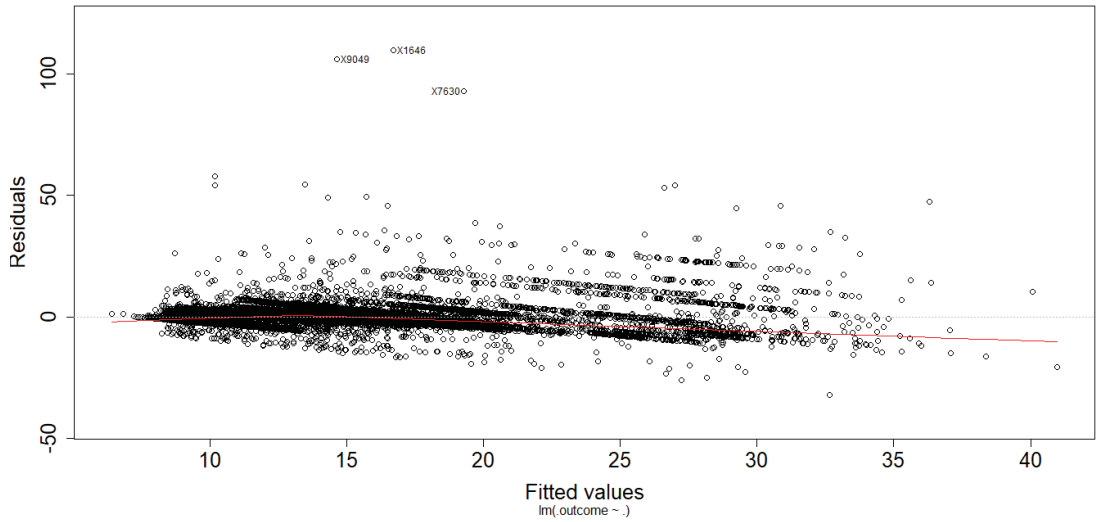
Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

ANEXO 2 – Análise dos pressupostos da Regressão Linear Múltipla

A média dos resíduos do modelo foi de -4.909979e-17, indicando que a média dos resíduos do modelo é próximo a zero.

A homocedasticidade dos resíduos foi analisada através da comparação dos resíduos e dos valores estimados, visto que há pouca alteração no resíduo conforme os valores estimados variam:

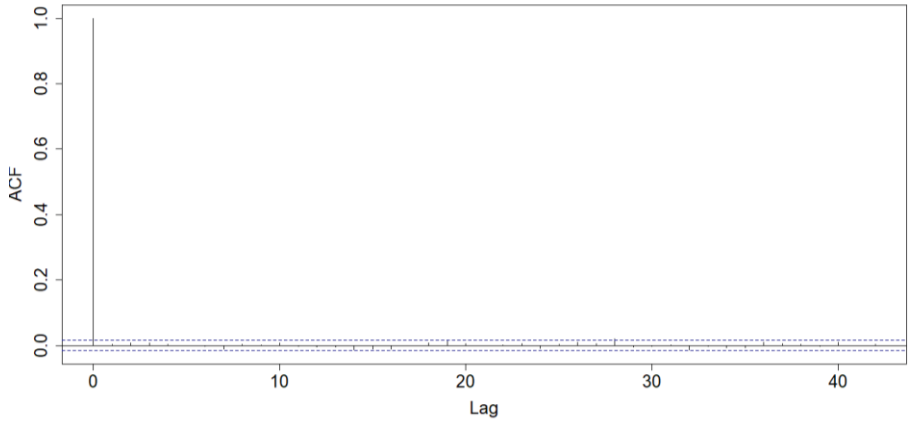
GRÁFICO 27: COMPARAÇÃO DE RESÍDUOS COM VALORES ESTIMADOS



Fonte: Elaborado pelo autor

Também foi realizado testes em relação a autocorrelação dos resíduos, sendo que no gráfico abaixo observa-se que não existe autocorrelação entre os resíduos para lags maiores que zero:

GRÁFICO 28: ANÁLISE DE AUTOCORRELAÇÃO



Fonte: Elaborado pelo autor

Foi utilizado VIF (Variance Inflation Factor) para analisar se existe problema de multicolinearidade, sendo que foi encontrado um valor de VIF baixo para a maioria dos preditores, com exceção de distância e estado do consumidor, indicando uma multicolonearidade baixa entre estas variáveis, como pode ser observado na tabela abaixo:

TABELA 13: VIF DE VARIÁVEIS NUMÉRICAS

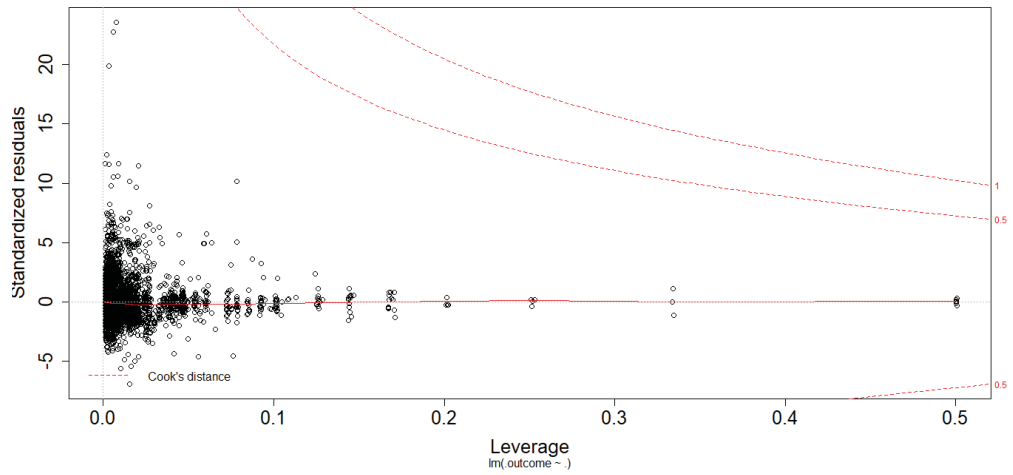
variable	VIF
price	1,141159
order_items_quantity	1,035499
product_category_name	2,772794
product_name_lenght	1,153313
product_description_lenght	1,244127
product_photos_qty	1,268544
customer_state	4,619905
distance	4,072298
trim	1,043074
hour	1,014813
day	1,010329
weekday	1,058538
weight_used	1,163942

Fonte: Elaborado pelo autor

Por fim, foi testado se existem valores influentes no modelo que alteram os parâmetros através da distância de Cook, sendo que nenhuma observação ultrapassou a distância de Cook de 0.5, indicando que não foram valores influentes

(indicando que a exclusão desses valores tende a não alterar o resultado da regressão) como pode ser observado no gráfico abaixo:

GRÁFICO 28: DISTÂNCIA DE COOK



Fonte: Elaborado pelo autor

ANEXO 3 – Resultados da Regressão Logística

term	estimate	std.error	statistic	p.value	significance
(Intercept)	1,71796	0,065277	26,31778	1,2E-152	***
payment_installments	-0,05497	0,004027	-13,6518	1,97E-42	***
price	0,00017	6,18E-05	2,721424	0,0065	**
freight_value	-0,00601	0,000633	-9,49657	2,17E-21	***
order_items_quantity	-0,44392	0,018588	-23,8825	4,7E-126	***
product_name_lenght	-0,00306	0,00105	-2,91617	0,003544	**
product_description_lenght	0,00011	1,68E-05	6,436991	1,22E-10	***
product_photos_qty	0,03338	0,006168	5,411733	6,24E-08	***
hour_survey_answer	0,00434	0,001419	3,05553	0,002247	**
disconfirmation	0,0729	0,001105	65,98069	0	***
payment_type_credit_card	0,02697	0,028351	0,951329	0,341437	
payment_type_debit_card	0,08492	0,089487	0,949013	0,342614	
payment_type_voucher	-0,13977	0,103042	-1,35641	0,17497	
used_voucher_yes	-0,16418	0,063746	-2,57548	0,01001	*

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

ANEXO 4 – Resultados de *t-test* dos parâmetros por tipo de pagamento

term	estimate	group	estimate.2	group.2	difference_means	p_value_test	test_significance
α	0,60471	credit_card	0,61058	boleto	-0,00587	0	***
β	0,374491	credit_card	0,30893	boleto	0,065562	0	***
λ	4,026735	credit_card	4,290621	boleto	-0,26389	0	***
α	0,622801	debit_card	0,61058	boleto	0,01222	0	***
β	0,353725	debit_card	0,30893	boleto	0,044796	8,79243E-11	***
λ	4,040522	debit_card	4,290621	boleto	-0,2501	1,90914E-17	***
α	0,598931	voucher	0,61058	boleto	-0,01165	0	***
β	0,328771	voucher	0,30893	boleto	0,019842	0,077297378	.
λ	3,705513	voucher	4,290621	boleto	-0,58511	8,15482E-16	***
α	0,622801	debit_card	0,60471	credit_card	0,018091	0	***
β	0,353725	debit_card	0,374491	credit_card	-0,02077	0,001019054	**
λ	4,040522	debit_card	4,026735	credit_card	0,013787	0,565364936	
α	0,598931	voucher	0,60471	credit_card	-0,00578	4,4288E-151	***
β	0,328771	voucher	0,374491	credit_card	-0,04572	0,000101378	***
λ	3,705513	voucher	4,026735	credit_card	-0,32122	2,08567E-07	***
α	0,598931	voucher	0,622801	debit_card	-0,02387	0	***
β	0,328771	voucher	0,353725	debit_card	-0,02495	0,050808149	.
λ	3,705513	voucher	4,040522	debit_card	-0,33501	2,80725E-07	***

Fonte: Elaborado pelo autor

*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

ANEXO 5 – Resultados de *t-test* dos parâmetros por Estado do consumidor

term	estimate	group	estimate.2	group.2	difference_means	p_value_test	test_significance
α	0,592568	PR	0,613602	SP	-0,02103	0	***
β	0,306405	PR	0,343151	SP	-0,03675	3,58E-27	***
λ	4,053962	PR	4,078055	SP	-0,02409	0,091467	.
α	0,588599	RJ	0,613602	SP	-0,025	0	***
β	0,370564	RJ	0,343151	SP	0,027413	0	***
λ	3,846497	RJ	4,078055	SP	-0,23156	0	***
α	0,583395	BA	0,613602	SP	-0,03021	0	***
β	0,340589	BA	0,343151	SP	-0,00256	0,107466	
λ	4,126827	BA	4,078055	SP	0,048772	1,84E-07	***
α	0,590475	MG	0,613602	SP	-0,02313	0	***
β	0,356768	MG	0,343151	SP	0,013617	3,51E-30	***
λ	4,002701	MG	4,078055	SP	-0,07535	2,5E-44	***
α	0,592568	PR	0,588599	RJ	0,003968	0	***
β	0,306405	PR	0,370564	RJ	-0,06416	4,58E-57	***
λ	4,053962	PR	3,846497	RJ	0,207465	3,15E-34	***
α	0,583395	BA	0,588599	RJ	-0,0052	0	***
β	0,340589	BA	0,370564	RJ	-0,02997	5,79E-57	***
λ	4,126827	BA	3,846497	RJ	0,28033	4,7E-109	***
α	0,590475	MG	0,588599	RJ	0,001875	0	***
β	0,356768	MG	0,370564	RJ	-0,0138	5,32E-28	***
λ	4,002701	MG	3,846497	RJ	0,156204	2,3E-127	***
α	0,592568	PR	0,583395	BA	0,009173	0	***

β	0,306405	PR	0,340589	BA	-0,03418	1,3E-21	***
λ	4,053962	PR	4,126827	BA	-0,07286	1,91E-05	***
α	0,590475	MG	0,583395	BA	0,00708	0	***
β	0,356768	MG	0,340589	BA	0,016179	1,21E-16	***
λ	4,002701	MG	4,126827	BA	-0,12413	3,92E-30	***
α	0,590475	MG	0,592568	PR	-0,00209	1,8E-292	***
β	0,356768	MG	0,306405	PR	0,050363	1,97E-41	***
λ	4,002701	MG	4,053962	PR	-0,05126	0,000706	***

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1

ANEXO 6 – Resultados de *t-test* dos parâmetros por categoria de produto

term	estimate	group	group 2	difference_means	p_value_test	test_significance
α	0,611961	beleza_saude	cama_mesa_banho	0,027759	0	***
β	0,372403	beleza_saude	cama_mesa_banho	0,016335	8,82E-36	***
λ	4,232992	beleza_saude	cama_mesa_banho	0,22089	1E-192	***
α	0,609204	esporte_lazer	cama_mesa_banho	0,025002	0	***
β	0,35786	esporte_lazer	cama_mesa_banho	0,001792	0,247136	
λ	4,087251	esporte_lazer	cama_mesa_banho	0,075149	7,02E-21	***
α	0,591869	informatica_acessorios	cama_mesa_banho	0,007667	0	***
β	0,3601	informatica_acessorios	cama_mesa_banho	0,004032	0,011097	*
λ	4,123197	informatica_acessorios	cama_mesa_banho	0,111095	8,28E-44	***
α	0,591395	moveis_decoracao	cama_mesa_banho	0,007193	0	***
β	0,331036	moveis_decoracao	cama_mesa_banho	-0,02503	8,17E-61	***
λ	3,921013	moveis_decoracao	cama_mesa_banho	-0,09109	1,65E-38	***
α	0,609204	esporte_lazer	beleza_saude	-0,00276	0	***
β	0,35786	esporte_lazer	beleza_saude	-0,01454	5,41E-20	***
λ	4,087251	esporte_lazer	beleza_saude	-0,14574	1,62E-63	***
α	0,591869	informatica_acessorios	beleza_saude	-0,02009	0	***
β	0,3601	informatica_acessorios	beleza_saude	-0,0123	2,81E-14	***
λ	4,123197	informatica_acessorios	beleza_saude	-0,1098	3,24E-41	***
α	0,591395	moveis_decoracao	beleza_saude	-0,02057	0	***
β	0,331036	moveis_decoracao	beleza_saude	-0,04137	5,6E-137	***
λ	3,921013	moveis_decoracao	beleza_saude	-0,31198	3,4E-255	***
α	0,591869	informatica_acessorios	esporte_lazer	-0,01733	0	***
β	0,3601	informatica_acessorios	esporte_lazer	0,002241	0,218179	
λ	4,123197	informatica_acessorios	esporte_lazer	0,035946	7,56E-05	***
α	0,591395	moveis_decoracao	esporte_lazer	-0,01781	0	***
β	0,331036	moveis_decoracao	esporte_lazer	-0,02682	4,86E-51	***
λ	3,921013	moveis_decoracao	esporte_lazer	-0,16624	1,01E-73	***
α	0,591395	moveis_decoracao	informatica_acessorios	-0,00047	6,73E-18	***
β	0,331036	moveis_decoracao	informatica_acessorios	-0,02906	3,22E-56	***
λ	3,921013	moveis_decoracao	informatica_acessorios	-0,20218	3,9E-105	***

Fonte: Elaborado pelo autor
*** p < 0.001; ** p < 0.01; * p <0.05; . p < 0.1